

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-202973

(43)Date of publication of application : 19.07.2002

(51)Int.Cl.

G06F 17/30

(21)Application number : 2001-291628

(71)Applicant : MATSUSHITA ELECTRIC IND CO
LTD

(22)Date of filing : 25.09.2001

(72)Inventor : SHIMOJIMA TAKASHI
ITO MASAO
TSURUBAYASHI TAKESHI
KATAYAMA OSAMU
NAKAI SHINICHI

(30)Priority

Priority number : 2000325286

Priority date : 25.10.2000

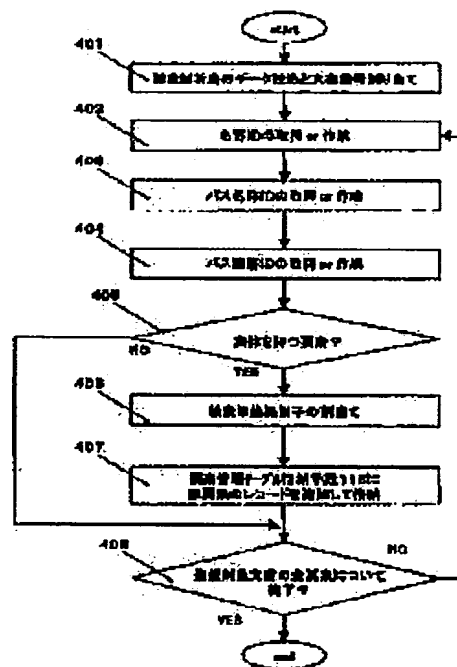
Priority country : JP

(54) STRUCTURED DOCUMENT MANAGEMENT DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a structured document management device capable of performing various retrievals designated in logical structure.

SOLUTION: In a document management system for handling structured documents, the information for specifying a logical structure position is managed with a pass name with tag names continuously described in order from the top hierarchy and a pass hierarchy with appearing orders of each hierarchy of the pass name continuously described, whereby various structured document retrievals can be realized.



LEGAL STATUS

[Date of request for examination]

05.10.2001

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2002-202973
(P2002-202973A)

(43) 公開日 平成14年7月19日 (2002.7.19)

(51) Int. Cl. ⁷	識別記号	F I	テームト (参考)
G 0 6 F 17/30	1 4 0	G 0 6 F 17/30	1 4 0 5 B 0 7 5
	4 1 9		4 1 9 A

審査請求 有 請求項の数26 O L (全 43 頁)

(21) 出願番号 特願2001-291628(P2001-291628)
(22) 出願日 平成13年9月25日 (2001.9.25)
(31) 優先権主張番号 特願2000-325286(P2000-325286)
(32) 優先日 平成12年10月25日 (2000.10.25)
(33) 優先権主張国 日本 (J P)

(71) 出願人 000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地
(72) 発明者 下島 崇
大阪府門真市大字門真1006番地 松下電器
産業株式会社内
(72) 発明者 伊藤 正雄
大阪府門真市大字門真1006番地 松下電器
産業株式会社内
(74) 代理人 100097445
弁理士 岩橋 文雄 (外2名)

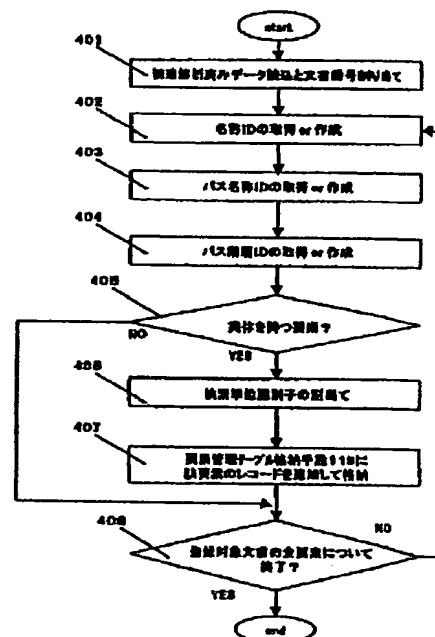
最終頁に続く

(54) 【発明の名称】 構造化文書管理装置

(57) 【要約】

【課題】 様々な論理構造を指定した検索をすることのできる構造化文書装置を提供する。

【解決手段】 構造化文書を扱う文書管理システムにおいて、論理構造位置を特定するための情報を、最上位階層から順にタグ名を連ねて記述したパス名称と、パス名称の各階層の出現順序を連ねて記述したパス階層で管理することにより、様々な構造化文書検索を実現することができる。



【特許請求の範囲】

【請求項1】 構造化文書を扱う文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体の前記木構造における位置を表現した要素実体位置識別子と、前記検索単位識別子から前記要素実体位置識別子を特定するために、少なくとも前記検索単位識別子と関係する前記要素実体位置識別子を対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件を入力する検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記要素実体位置識別子を特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する要素実体位置識別子を求め、前記要素実体位置識別子と前記検索条件解析手段により求めた前記要素実体位置識別子とが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項2】 構造化文書を扱う文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、前記検索単位識別子から前記パス名称IDと前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記パス名称ID、前記パス階層IDの少なくともいずれか1つを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応するパス名称ID又はパス階層IDを求め、前記パス名称ID又は前記パス階層IDと前記検索条件

解析手段により求めたパス名称ID又はパス階層IDとが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項3】 構造化文書を扱う構造化文書管理装置において、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により生成された木構造からタグ名を識別する名称IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件を入力する検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記名称IDを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称IDを求め、前記名称IDと前記検索条件解析手段で求めた名称IDとが一致する検索単位識別子のみを抽出する構造照合手段を備えた構造化文書管理装置。

【請求項4】 文字列検索結果一覧や各要素実体表示のためのデータを作成する結果作成手段と、前記結果作成手段で作成された検索結果を端末に表示する結果表示手段とを有することを特徴とする請求項1から3に記載の構造化文書管理装置。

【請求項5】 構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、前記検索単位識別子から前記パス名称ID及び前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成する構造情報作成手段とを備えた構造化文書登録装置。

【請求項6】 構造化文書の木構造が変化した場合に、要素管理テーブルに記録されたパス名称ID、パス階層IDのうち、変更が必要なIDを更新することを特徴とする請求項2記載の構造化文書管理装置。

【請求項7】 構造化文書の木構造が変化した場合に、要素管理テーブルに記録されたパス名称ID、パス階層IDのうち、変更が必要なIDを更新することを特徴と

する請求項5記載の構造化文書登録装置。

【請求項8】 構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により生成された木構造からタグ名を識別する名称IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成する構造情報作成手段とを備えた構造化文書登録装置。

【請求項9】 各要素実体内部にさらにタグに囲まれた要素実体（子要素）を含む構造化文書の索引作成において、各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成することを特徴とする文字列索引作成装置。

【請求項10】 予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の索引作成において、該タグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を作成する数値型索引作成手段を有していることを特徴とする請求項9記載の文字列索引作成装置。

【請求項11】 所定の条件に該当する文字列を検索する場合において、タグ名を識別する名称IDと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルまたは、前記検索単位識別子から前記パス名称IDと前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルの少なくともいずれか一方を記憶するデータ格納部と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件から検索条件に該当する前記名称ID、前記パス名称ID、前記パス階層IDの少なくともいずれか1つ（ID1）を特定する検索条件解析手段と、検索条件に該当する文字列を有する前記検索単位識別子を求める文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称ID、パス名称ID、パス階層IDの

少なくともいずれか1つ（ID2）を求め、前記ID2と前記検索条件解析手段により求めたID1とが一致する検索単位識別子のみを抽出する構造照合手段を備えた文字列検索装置。

【請求項12】 予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の数値範囲検索において、前記タグに囲まれた文字列を識別する独自の検索単位識別子と前記タグに囲まれた文字列を数値に変換した数値データとを対応付けた数値型索引を参照し、検索条件に該当する前記検索単位識別子を抽出する数値型索引検索手段を有することを特徴とする請求項11記載の文字列検索装置。

【請求項13】 木構造で表現した構造化文書の登録方法において、構造化文書を読み込むステップと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDを取得するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を取得するステップと、前記検索単位識別子から前記パス名称ID及び前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体。

【請求項14】 木構造で表現した構造化文書の登録方法において、構造化文書を読み込むステップと、タグ名を識別する名称IDを取得するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を取得するステップと、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体。

【請求項15】 要素実体内部にさらにタグに囲まれた要素実体（子要素）を有する構造化文書の文字索引の生成方法について、構造解析済みデータを読み込むステップと、要素実体を有するか否かをチェックするステップと、要素実体を識別するための検索単位識別子を取得するステップと、前記子要素を含むか否かを調べるステップと、該子要素を識別する検索単位識別子を取得するステップと、要素実体から1以上の所定文字数を単位とする文字列を取り出すステップと、前記文字列の各文字の属する検索単位識別子を求めるステップと、該文字列及び該文字列の各文字の属する前記検索単位識別子及び前記タグを取り除いた要素実体内での当該文字列の位置を示す文字位置識別子を有する検索文字列索引を生成するステップとを有するプログラムを記録した可搬型媒体。

【請求項16】 構造化文書の数値検索用索引生成方法について、構造解析済みデータを読み込むステップと、

予め数値であることを定義しているタグに囲まれた文字列であるか否かを判断するステップと、数値であることを定義したタグに囲まれた文字列を識別するための検索単位識別子を取得するステップと、該文字列を数値に変換するステップと、前記検索単位識別子と前記変換された数値とからなる数値型索引を生成するステップを有するプログラムを記録した可搬型媒体。

【請求項 17】 構造化文書の検索方法について、検索条件を読み込むステップと、前記検索条件に該当するタグ名を識別する名称 ID 又は、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称 ID 又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層 ID の少なくともいずれか 1 つの ID (以下、ID 1) に変換するステップと、検索条件に該当する文字列を有する各要素実体を識別する検索単位識別子 (以下、ID 2) を特定するステップと、前記 ID 2 から前記名称 ID、前記パス名称 ID、前記パス階層 ID を特定するために、少なくとも前記 ID 2 と関係する前記名称 ID、前記パス名称 ID、前記パス階層 ID を対応付けた要素管理テーブルを参照し、前記 ID 2 に対応する前記名称 ID、前記パス名称 ID、前記パス階層 ID の少なくともいずれか 1 つの ID (以下、ID 3) を求めるステップと、前記 ID 1 と前記 ID 3 とが一致する前記検索単位識別子のみを抽出するステップを有するプログラムを記録した可搬型媒体。

【請求項 18】 中間ノード以下を検索範囲に指定した場合における検索範囲に含まれるノードを決定する方法について、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を、1 階層登り、現在位置するノードが指定した中間ノードと一致するか又は、既に検索範囲に含まれていると判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲に含まれると判定し、現在位置するノードが指定した中間ノードと一致しないか又は、既に検索範囲外と判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲外であると判定する処理を、最下層ノードを起点として 1 階層登る毎に実行し、最上位層のノードに至るまで繰り返し実行することにより検索範囲を特定する方法。

【請求項 19】 構造化文書を管理するために汎用計算機を、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称

を識別するパス名称 ID と、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層 ID と、前記検索単位識別子から前記パス名称 ID と前記パス階層 ID を特定するために、少なくとも前記検索単位識別子と関係する前記パス名称 ID 及びパス階層 ID を対応付けた要素管理テーブルを作成する構造情報作成手段と、文字列検索を行うための文字列索引を作成する文字列索引作成手段と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件に該当する前記パス名称 ID、前記パス階層 ID の少なくともいずれか 1 つを特定する検索条件解析手段と、前記文字列索引作成手段で作成された文字列索引を用いて検索条件に該当する文字列を有する各要素実体の前記検索単位識別子を特定する文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応するパス名称 ID 又はパス階層 ID を求め、前記パス名称 ID 又は前記パス階層 ID と前記検索条件解析手段により求めたパス名称 ID 又はパス階層 ID とが一致する検索単位識別子のみを抽出する構造照合手段、として機能させるための構造化文書管理プログラム。

【請求項 20】 各要素実体内部にさらにタグに囲まれた要素実体 (子要素) を含む構造化文書の索引を作成するために汎用計算機を、各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成する文字列索引作成手段として機能させるための文字索引作成プログラム。

【請求項 21】 予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の索引を作成するために汎用計算機を、該タグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を作成する数値型索引作成手段として機能させるための文字列索引作成プログラム。

【請求項 22】 木構造で表現されるデータにおいて所定のノード以下を検索範囲に指定した場合に、検索範囲に含まれるノードを特定するプログラムであって、各ノードが検索範囲に含まれるか否かを示す照合フラグを格納する照合テーブルを初期化する第一のステップ、参照しているノードが検索範囲内か否か又は未定であるかを、照合テーブルをもとに判断する第二のステップ、第二のステップにより検索範囲内と判断した場合は、参照しているノードについて検索範囲内を示す照合フラグを照合テーブルに設定する第三のステップ、第二のステッ

ブにより検索範囲外と判断した場合は、参照しているノードについて検索範囲外を示す照合フラグを照合テーブルに設定する第四のステップ、第二のステップにより未定と判断した場合であって、さらに参照しているノードが指定したノードと一致する場合又は、既に検索範囲内である場合は、それまでたどった全てのノードについて検索範囲内を示す照合フラグを照合テーブルに設定する第五のステップ、第二のステップにより未定と判断した場合であって、さらに参照しているノードが既に検索範囲外である場合は、それまでたどった全てのノードについて範囲外を示す照合フラグを照合テーブルに設定する第六のステップ、第五のステップまたは第六のステップのいずれにも該当しない場合は、現在参照しているノードから1階層上る第七のステップ、前記第七のステップにより1階層上ったノードがルートノードである場合は、それまでたどった全てのノードについて検索範囲外を示す照合フラグを照合テーブルに設定する第八のステップ、前記第七のステップにより1階層上ったノードがルートノード以外である場合は、前記第五のステップへ戻る第八のステップ、とから構成されることにより、検索範囲を特定するプログラム。

【請求項23】 木構造で表現される構造化文書を管理する装置であって、要素実体を識別する検索単位識別子を割当てる構造情報作成手段と、前記検索単位識別子とは別個に要素実体を特定する手段として、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段と、前記木構造においてタグ名を階層別に連ねたパス名称を格納する手段と、を備え、さらに、前記パス階層及びパス名称と前記検索単位識別子とを関連付ける要素管理テーブルを格納する手段と、検索条件の文字列を含む要素実体の検索単位識別子を抽出する文字列索引検索手段と、文字列索引検索手段により抽出された検索単位識別子から、前記要素管理テーブルを参照し、検索条件として指定したパス階層又はパス名称を満たす文書を検索する構造照合手段と、を有する構造化文書管理装置。

【請求項24】 木構造で表現可能なデータ構造を有するデータを管理するデータ管理装置であって、データの実体要素の特定は、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段を用いることを特徴とするデータ管理装置。

【請求項25】 木構造で表現されたデータのタグ名を階層別に連ねたパス名称を格納する手段をさらに備え、前記木構造におけるデータの実体要素を一意に特定するために前記パス階層を格納する手段と、前記パス名称を格納する手段とを用いることを特徴とする請求項24記載のデータ管理装置。

【請求項26】 同一親ノードを持ち同一のタグ名称を有する実体要素が複数存在する場合、前記パス名称は同

一に表現されることを特徴とする請求項25記載のデータ管理装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、SGMLやXMLなどの論理的な構造要素を有する構造化文書を計算機を用いて管理する文書管理システムにおける、論理構造を指定した検索を行なう構造化文書検索方法に関するものである。

【0002】

【従来の技術】電子化文書の増大に伴い、マニュアル、議事録、仕様書等、論理的構造を有する文書を扱う構造化文書に対する関心が高まっている。それにより、文書内容のみによる検索だけでなく、構造化文書の特長を生かした、論理構造を指定した検索を行なう機能が重要となる。構造化文書はその論理構造がDTD(Document Type Definition:文書型定義)によって定義される。

【0003】従来、構造化文書管理システムにおける文書の検索装置としては、特開平10-240752号公報(以下、公知例と呼ぶ)に記載された発明が知られている。

【0004】以下、公知例の概要について説明する。その文書登録システムの構成図は図33に示すとおりである。公知例では登録する際、まず文書構造解析プログラム3301にて登録対象文書の持つ論理構造を解析して、解析済み文書データを作成し解析済み文書データ格納領域3305に登録する。

【0005】次に、構造インデックス作成プログラム3302にて各登録対象文書の持つ論理構造を、登録順に従って順次重ね合わせ、文書中における出現位置および種別が同じである要素群は単一のメタ要素によって代表させ、文書中における出現位置が同じである文字列データ群は単一のメタ文字列データによって代表させることにより、メタ要素群およびメタ文字列データ群(公知例ではこれらを総称してメタノードと呼ぶ)の木構造から構成される構造インデックスを生成し該構造インデックスを構成する全てのメタノードに対して、それらを構造インデックスの中で一意に識別する識別子(公知例ではこれを文脈識別子と呼ぶ)を与え、構造インデックス格納領域3306に登録する。

【0006】図34は上記構造インデックスを作成する過程を示す図である。図34において、文書1、文書2、文書3は、それぞれ登録対象文書の解析済み文書データを表わしている。これらの解析済み文書データの構造を既存の構造インデックス上に順次重ね合わせることで、構造インデックスが形成されていく。まず最初に文書1が入力されると、最初の段階では構造インデックスは初期状態(空)であるため、該解析済みデータと等価な木構造が生成されてそのまま構造インデックスに

登録され、構造インデックスは3401に示す状態となる。新たに生成されたメタ要素にはE1からE5までの文脈識別子、新たに生成されたメタ文字列データにはC1からC3までの文脈識別子が割り当てられる。次に文書2が入力されると、既存の構造インデックス(3401)と構造が重複する部分については何も行わず、3401上に対応する部分がなかった部分構造(図中の網掛け部分)だけが新たに登録される。新たに生成されたメタ要素には文脈識別子E6およびE7、新たに生成されたメタ文字列データには文脈識別子C4が割り当てられる。次に文書3が入力されると、既存の構造インデックス(3402)と構造が重複する部分については何も行わず、3402上に対応する部分がなかった部分構造(図中の網掛け部分)だけが新たに登録される。新たに生成されたメタ要素には文脈識別子E8、E9およびE10、新たに生成されたメタ文字列データには文脈識別子C5およびC6が割り当てられる。このようにして、3個の文書が登録された段階で、構造インデックスは3403に示す状態となる。

【0007】次に、構造化全文データ生成プログラム3303にて各登録対象文書について、その文書に対応する解析済み文書データ中に含まれるすべての文字列と、その文字列を構造インデックス中で示される文脈識別子との対応関係の定義から構成されるデータ(公知例ではこれを構造化全文データと呼ぶ)を生成し、構造化全文データ格納領域3307に登録する。

【0008】次に、文字列インデックス作成プログラム3304にて、各登録対象文書に対応する構造化全文データから、前記文脈識別子を含んだ全文検索を行なうための文字列インデックスを作成し、文字列インデックス格納領域3308に登録する。

【0009】図35は、文字列インデックスの例を示したものであり、部分文字列(3404)を2文字とした場合の例を示している。各部分文字列に対して該部分文字列を含む文書を識別する文書識別子(3405)、該文書中において前記部分文字列を含む文字列データの文書構造中における位置を識別する文脈識別子(3406)、文書中における前記部分文字列の文字位置(3407)から構成されている。なお、図中の“X”は文字列の直前に位置する文字の位置を“X”として相対的な文字位置を示している。

【0010】また、公知例における検索は、まず前記構造インデックスを参照し、指定された構造条件を満たす文脈識別子の集合を決定する。

【0011】次に、それらの文脈識別子をキーとして文字列を検索することにより、指定条件を満たす文書群を求める。

【0012】また、公知例における登録の際に、例えば強調表示のような非構造的要素(Mixed Contentと呼ぶ：詳細は実施の形態3で説明する)が含ま

れる場合、該構造を無視して文字列インデックスを作成する。

【0013】

【発明が解決しようとする課題】上記従来技術の方法では、図35に示すように全文検索を行なうための文字列インデックス内に、登録文書を識別する文書識別子と、論理構造に関する情報である文脈識別子と、文字連鎖の位置を示す文字位置という3要素のデータを含んでいるため、前記文字列インデックスのサイズが大きくなり、そのためメモリ量が増大し、装置のコストアップにつながるという課題を有していた。

【0014】また上記従来技術の方法では、図35に示すように文字列インデックス内の各文字連鎖に論理構造に関する情報である文脈識別子を含んでいるため、複数の登録文書の1つについて要素実体を追加、変更したことにより、複数の登録文書の論理構造を順次重ね合わせることで形成される構造インデックス(図34)が変化した場合、文字列インデックスの文脈識別子を更新する必要が発生し、要素実体の文字連鎖数が膨大の場合、処理量も膨大になるという課題を有していた。

【0015】以下、この課題について具体例を通して詳細に説明する。

【0016】図36は2つの文書が登録されている場合の例で、このうち1つの登録文書を変更する例を示している。この例では、文書1と文書2の論理構造は同一であるので、作成される構造インデックスの論理構造も文書1又は2と同一である。この例では、文書2の第1章と第2章の間に新たな章を追加して3つの章から成る文書に変更する例を示している。すなわち文書2に新たに第2章となるブロック(図36の4000)を追加する例を示している。このとき、変更前に第2章であったブロック(図36の4001)が第3章となるが、変更前の構造インデックスには、文書1、2とも第2章までしかなかった為、文書2の第3章に相当する文脈識別子は存在しない(図36の変更前構造インデックス)。そこで、図36(変更後構造インデックス)に示すように構造インデックスを更新する必要がある。

【0017】図36の更新後の構造インデックスに示すように、文書2で新たに第3章となった要素実体に対応する文脈識別子は‘C4’となっている。しかし変更前、前記要素実体に対応する文脈識別子は‘C3’であったので、前記要素実体の文字列インデックスに保持されている各文字連鎖の文脈識別子を‘C3’から‘C4’に変更する必要がある。例えば、文書2の第3章に相当する要素実体が100文字から構成されているとすると、2文字連鎖で索引を作成する場合、99個の文字連鎖について文脈識別子を変更する必要がある。このように要素実体の文字連鎖数に応じて処理量も大きくなってしまいう課題を有していた。

【0018】なお、変更後に第2章となった要素実体に

新たな文脈識別子‘C4’を付与し、変更前第2章で変更後第3章となる要素実体にはそのままの文脈識別子

‘C3’とする更新方法も考えられるが、この場合は文書1の第2章に相当する要素実体の文字列インデックスの文字連鎖について、文脈識別子を‘C3’から‘C4’に変更する必要がある。この例では登録文書が2つなので、上述の方法と更新にかかる処理量は同一であるが、登録文書の数が増加した場合、第2章を有する全ての登録文書の要素実体について、その文字列インデックスを文脈識別子を‘C3’から‘C4’に変更する必要があるため、かえって処理量が増加してしまう結果になる。

【0019】また別の課題として、従来技術の構造インデックスは図34に示すように複数の登録文書の論理構造を順次重ね合わせることによって形成されるので、登録文書の論理構造がほぼ同一の場合は新たに文脈識別子を付与する機会は少ないが、各登録文書の論理構造が大きく異なる場合は論理構造の重なりが少なくなり、このような論理構造が異なる登録文書が膨大に登録された場合は、文脈識別子の数が膨大になるという課題を有していた。

【0020】また従来技術の構造インデックスは、図34に示すように複数の登録文書の論理構造を順次重ね合わせるによって形成されるので、この方法により形成される構造インデックスには、1つの親ノードから同一のタグ名を有する子ノードが複数出ている構造も発生する場合がある。このとき検索範囲として或るタグ名を指定した場合、各ノードのタグ名が該当するタグ名であるか否かをチェックする必要があるが、たとえ上記のように1つの親ノードから同一のタグ名を有する子ノードが複数出ていたとしても、各子ノードの1つ1つについて該当するタグ名を有するノードであるか否かをチェックするOR検索が必要の為、検索が遅くなるという課題を有していた。

【0021】また上記従来方法では、要素実体である“段落”要素中にMixed Contentとして“キーワード”要素を含むような場合、“キーワード”タグの構造を無視して文字列インデックスを作成するため、“キーワード”タグの中に“〇〇”を含む文書”というような検索条件に対応できないという課題を有していた。

【0022】本発明は上記従来技術の課題を解決するもので、構造化文書を対象とした全文検索において、様々な論理構造指定検索に対応すること、さらに検索用索引のサイズ削減、文書の一部変更・一部削除時における検索用索引の変更作業の簡易化、中間ノード以下を指定した高速な検索、そしてMixed Contentにまたがる検索、およびMixed Contentである要素を指定した検索を行なうことを目的とする。

【0023】

【課題を解決するための手段】上記課題を解決するため

に、請求項1では各要素実体を識別する検索単位識別子と、各要素実体の前記木構造における位置を表現した要素実体位置識別子と、前記検索単位識別子から前記要素実体位置識別子を特定するために、少なくとも前記検索単位識別子と関係する前記要素実体位置識別子を対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように文書構造が変化する度に文字列インデックス内の文脈識別子を変更する必要はないので、登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。

【0024】請求項2では各要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、前記検索単位識別子から前記パス名称IDと前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。また、パス名称IDとパス階層IDを導入することにより、従来技術のように検索範囲を特定する際のOR検索が不要になる。また、登録文書の論理構造が異なる文書を多く登録する場合でも、要素実体をパス名称IDとパス階層IDとで特定するので、従来技術のように複数の登録文書の論理構造を順次重ね合わせるによって形成される場合に必要となる文脈識別子数よりは少なく済む。

【0025】請求項3ではタグ名を識別する名称IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成する構造情報作成手段を有することにより、検索範囲として登録文書のノードのタグ名を指定することが可能となる。

【0026】請求項4では文字列検索結果一覧や各要素実体表示のためのデータを作成する結果作成手段と、前記結果作成手段で作成された検索結果を端末に表示する結果表示手段とを有することにより、使用者に検索結果を表示することが可能となる。

【0027】請求項5ではネットワーク上に、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の木構造を生成する構造解析手段と、前記構造解析手段により木構造で表現された構造化文書において、各

要素実体を識別する検索単位識別子と、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、前記検索単位識別子から前記パス名称ID及び前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成する構造情報作成手段とから成る構造化文書登録部を独立して設けることにより、ネットワークを介して遠隔から構造化文書の登録をすることが可能となり、登録文書の構造が変化した場合でも前記要素管理テーブルを更新するのみで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。またパス名称IDとパス階層IDを導入することにより、従来技術のように検索範囲を特定する際のOR検索が不要になる。また登録文書の論理構造が異なる文書を多く登録する場合でも、要素実体をパス名称IDとパス階層IDとで特定するので、従来技術のように複数の登録文書の論理構造を順次重ね合わせることによって形成される場合に必要となる文脈識別子数よりは少なくて済む。

【0028】請求項6及び7では構造化文書の本構造が変化した場合に、要素管理テーブルに記録されたパス名称ID、パス階層IDのうち、変更が必要なIDを更新することにより、登録文書の構造が変化した場合でも前記要素管理テーブルを更新することで対応が可能となり、従来技術のように登録文書の論理構造の変化する度に文字列インデックス更新のための膨大な処理量が発生することはない。

【0029】請求項8ではネットワーク上に、構造化文書の入力を行う構造化文書入力手段と、前記構造化文書入力手段により取り込んだ構造化文書を解析し該構造化文書の本構造を生成する構造解析手段と、前記構造解析手段により生成された本構造からタグ名を識別する名称IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成する構造情報作成手段とから成る構造化文書登録部を独立して設けることにより、ネットワークを介して遠隔から構造化文書の登録が可能となる。

【0030】請求項9では各要素実体から所定の文字数で取り出した文字列が前記タグにまたがる場合は、該子要素を識別する独自の検索単位識別子を取得し、該文字列と該文字列の各文字の属する要素実体を識別する検索単位識別子と前記タグを取り除いた要素実体内での該文字列の位置を示す文字位置識別子とから成る検索用文字列索引を生成する文字列索引作成部により、Mixed Contentを含んだ構造化文書でも検索が可能と

なる。また作成される文字索引は前記検索単位識別子と前記文字位置識別子の2要素から成るので、従来技術では3要素から成る文字列インデックスと比べメモリ量を削減することができ、装置のコストダウンを実現することができる。

【0031】請求項10では予め数値であることを定義しているタグに囲まれた文字列を識別する独自の検索単位識別子を取得し、該タグに囲まれた文字列を数値データに変換し、前記検索単位識別子と前記数値データとを対応付けた数値型索引を生成する数値型索引作成手段により、特定の数値範囲を指定した検索が可能になる。

【0032】請求項11ではネットワーク上に、タグ名を識別する名称IDと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDと、各要素実体を識別する検索単位識別子と、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルまたは、前記検索単位識別子から前記パス名称IDと前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルの少なくともいずれか一方を記憶するデータ格納部と、検索条件の入力を行う検索条件入力手段と、前記検索条件入力手段で入力された検索条件から検索条件に該当する前記名称ID、前記パス名称ID、前記パス階層IDの少なくともいずれか1つ（ID1）を特定する検索条件解析手段と、検索条件に該当する文字列を有する前記検索単位識別子を求める文字列索引検索手段と、前記文字列索引検索手段で特定した検索単位識別子を基に前記要素管理テーブルを参照して対応する名称ID、パス名称ID、パス階層IDの少なくともいずれか1つ（ID2）を求め、前記ID2と前記検索条件解析手段により求めたID1とが一致する検索単位識別子のみを抽出する構造照合手段を備えた文字列検索部をそれぞれ独立して設けることにより、ネットワークを介して遠隔からの文字列検索が可能となる。

【0033】請求項12では予め数値であることを定義しているタグに囲まれた文字列を含む構造化文書の数値範囲検索において、前記タグに囲まれた文字列を識別する独自の検索単位識別子と前記タグに囲まれた文字列を数値に変換した数値データとを対応付けた数値型索引を参照し、検索条件に該当する前記検索単位識別子を抽出する数値型索引検索手段を有することを特徴とする請求項11記載の文字列検索部を有していることにより、ネットワークを介して遠隔から、指定した範囲の数値を有する要素実体の検索単位識別子を求めることが可能となる。

【0034】請求項13では構造化文書を読み込むステ

ップと、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称IDと、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDを生成するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を生成するステップと、前記検索識別子から前記パス名称ID及び前記パス階層IDを特定するために、少なくとも前記検索単位識別子と関係する前記パス名称ID及びパス階層IDを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることで、構造化文書を登録する構造化文書登録部の機能を持たせることが可能となる。

【0035】請求項14では構造化文書を読み込むステップと、タグ名を識別する名称IDを生成するステップと、要素実体を有するか否かを判断するステップと、各要素実体を識別する検索単位識別子を生成するステップと、前記検索単位識別子から前記名称IDを特定するために、少なくとも前記検索単位識別子と関係する前記名称IDを対応付けた要素管理テーブルを作成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることで、構造化文書を登録する構造化文書登録部の機能を持たせることが可能となる。

【0036】請求項15では、要素実体内部にさらにタグに囲まれた要素実体（子要素）を有する構造化文書の文字索引の生成方法について、構造解析済みデータを読み込むステップと、要素実体を有するか否かをチェックするステップと、要素実体を識別するための検索単位識別子を取得するステップと、前記子要素を含むか否かを調べるステップと、該子要素を識別する検索単位識別子を取得するステップと、要素実体から1以上の所定文字数を単位とする文字列を取り出すステップと、前記文字列の各文字の属する検索単位識別子を求めるステップと、該文字列及び該文字列の各文字の属する前記検索単位識別子及びタグを取り除いた要素実体内での当該文字列の位置を示す文字位置識別子を有する検索文字列索引を生成するステップとを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、Mixed Contentを含んだ構造化文書でも検索が可能な文字列索引を作成する文字列索引作成部の機能を持たせることが可能となる。

【0037】請求項16では、構造化文書の数値検索用索引生成方法について、構造化文書を読み込むステップと、予め数値であることを定義しているタグに囲まれた文字列であるか否かを判断するステップと、数値であることを定義したタグに囲まれた文字列を識別するための検索単位識別子を取得するステップと、該文字列を数値

に変換するステップと、前記検索単位識別子と前記数値とからなる数値型索引を生成するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、数値範囲を指定した検索も可能な文字列索引を生成する文字列索引作成部の機能を持たせることが可能となる。

【0038】請求項17では、構造化文書の検索方法について、検索条件を読み込むステップと、前記検索条件に該当するタグ名を識別する名称ID又は、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称ID又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を識別するパス階層IDのいずれかのID（以下、ID1）に変換するステップと、検索条件に該当する文字列を有する各要素実体を識別する検索単位識別子（以下、ID2）を特定するステップと、前記ID2から前記名称ID、前記パス名称ID、前記パス階層IDを特定するために、少なくとも前記ID2と関係する前記名称ID、前記パス名称ID、前記パス階層IDを対応付けた要素管理テーブルを参照し、前記ID2に対応する前記名称ID、前記パス名称ID、前記パス階層IDの少なくともいずれか1つのID（以下、ID3）を求めるステップと、前記ID1と前記ID3とが一致する前記検索単位識別子のみを抽出するステップを有するプログラムを記録した可搬型媒体により、汎用計算機に上記プログラムをインストールすることにより、文字列検索部の機能を持たせることが可能となる。

【0039】請求項18では、中間ノード以下を検索範囲に指定した場合における検索範囲に含まれるノードを決定する方法について、各要素実体に至るタグ名を階層順に連ねたパス名称を識別するパス名称又は、同一の親ノードを持ち同一な名称を持つタグの同一階層内での出現順序を階層順に連ねたパス階層を、1階層登り、現在位置するノードが指定した中間ノードと一致するか又は、既に検索範囲に含まれていると判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲に含まれると判定し、現在位置するノードが指定した中間ノードと一致しないか又は、既に検索範囲外と判定されているノードであるかを判断し、前記いずれかの条件に該当するノードである場合はそれまでたどったノード全てを検索範囲外であると判定する処理を、最下層ノードを起点として1階層登る毎に実行し、最上位層のノードに至るまで繰り返し実行することにより検索範囲を特定する方法により、検索範囲として或る中間ノード以下を指定した場合に検索範囲に含まれるノードを特定することが可能となる。

【0040】請求項19の発明は、請求項2に記載の構造化文書管理装置を、汎用計算機とプログラムによって実現することを可能とするものである。

【0041】請求項20の発明は、請求項9に記載の文字列索引作成装置を、汎用計算機とプログラムによって実現することを可能とするものである。

【0042】請求項21の発明は、請求項10に記載の文字列索引作成装置を、汎用計算機とプログラムによって実現することを可能とするものである。

【0043】請求項22の発明は、木構造を有するデータを検索するために、検索範囲として所定のノード以下を指定した場合に、各ノードが検索範囲に含まれるか否かを示す照合テーブルを順次作成していくプログラムにより、検索範囲の特定を効率良く実現するものである。

【0044】請求項23の発明は、木構造で表現される構造化文書を管理する装置であって、要素実体を識別する検索単位識別子を割当てる構造情報作成手段と、前記検索単位識別子とは別個に要素実体を特定する手段として、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段と、前記木構造においてタグ名を階層別に連ねたパス名称を格納する手段と、を備え、さらに、前記パス階層及びパス名称と前記検索単位識別子とを関連付ける要素管理テーブルを格納する手段と、検索条件の文字列を含む要素実体の検索単位識別子を抽出する文字列索引検索手段と、文字列索引検索手段により抽出された検索単位識別子から、前記要素管理テーブルを参照し、検索条件として指定したパス階層又はパス名称を満たす文書を検索する構造照合手段と、を有する構造化文書管理装置であって、効率良く文書検索を実現することが可能となる。

【0045】請求項24の発明は、木構造で表現可能なデータ構造を有するデータを管理するデータ管理装置であって、データの実体要素の特定は、前記木構造において同一の親ノードを持ち同一な名称を持つタグの出現順序を階層別に連ねたパス階層を格納する手段を用いることを特徴とするデータ管理装置であって、木構造で表現できるデータの管理を少ない個数のIDにより管理することが可能になる。

【0046】請求項25の発明は、木構造で表現されたデータのタグ名を階層別に連ねたパス名称を格納する手段をさらに備え、前記木構造におけるデータの実体要素を一意に特定するために前記パス階層を格納する手段と、前記パス名称を格納する手段とを用いることを特徴とする請求項24記載のデータ管理装置であって、木構造で表現できるデータをパス階層及びパス名称で特定することにより少ない個数のIDにより管理することが可能になる。

【0047】請求項26の発明は、同一親ノードを持ち同一のタグ名称を有する実体要素が複数存在する場合、前記パス名称は同一に表現されることを特徴とする請求項25記載のデータ管理装置であって、データの検索においていわゆるOR検索が不要となり、高速に検索する

ことが可能にすることが可能となる。

【0048】

【発明の実施の形態】以下、本発明の実施の形態について説明する。なお、本発明はこれら実施の形態に何ら限定されるものではなく、その要旨を逸脱しない範囲において、種々なる態様で実施し得る。

【0049】（実施の形態1）図1は本発明の実施の形態1における構造化文書管理装置の構成図である。図1に示す構造化文書管理装置は、端末101、構造化文書入力手段102、検索条件入力手段103、結果表示手段104、検索エンジン105、データ格納部106からなる。

【0050】端末101は、文書検索における検索条件の指定および検索結果の表示に使用する。

【0051】構造化文書入力手段102は、登録対象文書を格納しておき、文書の登録を行なう際にここから検索エンジン105へデータを送る。検索条件入力手段103は、端末101から入力された検索条件を検索エンジン105へ送る。

【0052】結果表示手段104は、検索結果を検索エンジン105から受け取り、端末101に表示する。

【0053】検索エンジン105は、実際に構造化文書の登録、検索および検索結果の作成を行なう。まず、登録に関して、107は登録対象文書の論理構造を解析する構造解析手段、108は前記構造解析手段によって論理構造に分けられた各要素の論理構造に関する情報を作成する構造情報作成手段、109は文字列に対して高速に検索を行なうための文字列索引を作成する文字列索引作成手段である。これら107、108、および109についての詳細は、文書登録処理の流れの説明の中で述べる。次に検索に関して、110は検索条件入力手段103から受けた検索条件中の論理構造に関する条件を、本検索エンジン内における構造条件の表現方法に変換する検索条件解析手段、111は前記文字列索引を用いて検索条件中の検索文字列で検索処理を行なう文字列索引検索手段、112は前記文字列索引検索手段で得られた文字列検索結果群の中から、前記検索条件解析手段で変換した本検索エンジン内における構造条件に一致するものだけを抽出する構造照合手段である。110、111、および112についての詳細は文書検索の流れの説明の中で述べる。次に結果作成に関して、113は検索結果の一覧や、実体表示のためのデータを作成し結果表示手段104へ渡す結果作成手段である。

【0054】データ格納部106は、構造解析手段107によって作成された構造解析済みデータを格納する構造解析済みデータ格納手段114、文書中の検索対象要素ごとに論理構造情報を格納した要素管理テーブル格納手段115、最上位階層から順にタグ名を連ねて記述した文字列（以下、パス名称と呼ぶ）を管理し、各パス名称にIDを割当てたパス名称インデックスを格納するパ

ス名称インデックス格納手段116、パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素かを示す番号）を連ねて記述した文字列（以下、パス階層と呼ぶ）を管理し、各パス階層にIDを割当てたパス階層インデックスを格納するパス階層インデックス格納手段117、各要素のタグ名に対してIDを割当てた名称IDテーブルを格納する名称IDテーブル格納手段118、前記文字列索引作成手段109によって作成された文字列索引を格納する文字列索引格納手段119、登録文書の実体データを格納する実体データ格納手段120、検索結果一覧のためのデータを格納する一覧データ格納手段121からなり、構造化文書の検索および結果表示に用いるデータの格納に使用する。

【0055】次に本実施の形態における文書登録の処理を具体的な構造化文書の例を用いて説明する。

【0056】まず、構造化文書入力手段102から登録対象文書を読み込む。次に構造解析手段107によって登録対象文書の構造を理解できる形に変換する。この構造解析手段107によって、文字の並びとしての構造化文書が構造情報作成手段108に理解できるデータ構造に変換され（以下、構造解析済みデータと呼ぶ）、構造解析済みデータ格納手段114に格納される。

【0057】次に構造情報作成手段108で、前記構造解析手段によって論理構造に分けられた各要素の論理構造に関する情報を作成する。

【0058】図2は構造化文書の一例である。図2の構造化文書を構造解析手段107によって解析した結果得られる木構造は図3のようになる。以下ではこの論理構造を持った構造化文書を中心に説明する。図3において実体（テキスト）を持つ要素（以下、要素実体）は網掛けで表示されている。またこれら要素実体は、検索エンジン内で検索単位を一意に表す符号（以下、検索単位識別子と呼ぶ）が割当てられる。この検索単位識別子は対象とする文書内の論理的な位置とは無関係な符号であり、例えば、数値であっても良い。

【0059】図3において要素実体の下段に書かれた数値が検索単位識別子の例である。また、要素実体は上述のパス名称インデックス、パス階層インデックス、名称IDのいずれか1つ又は上記インデックスの組み合わせにより特定が可能なので、上記3種のインデックスを総称して「要素実体位置識別子」という。

【0060】図4は構造情報作成手段108の処理の流れである。まず、登録対象文書の構造解析済みデータを構造解析済みデータ格納手段114から読み込み、登録対象文書ごとに一意な番号（以下、文書番号と呼ぶ）を割当てる（ステップ401）。

【0061】次に登録対象文書の各要素に対して以下の処理を繰り返す。まず、現在参照中の要素の名称IDの取得を行なう（ステップ402）。図5は図3のような

木構造を持つ構造化文書を登録した時に最終的に作成される名称IDテーブルの例である。図3の301の要素のタグ名は「段落」であるから、図5より名称IDは「T9」である。ステップ402では、この名称IDテーブルに現在参照中の要素に該当するタグ名と名称IDのレコードが存在する場合はその名称IDを取得し、存在しない場合にはそのタグ名と名称IDのレコードを新たに作成し、名称IDテーブル格納手段118に格納するとともに、その名称IDを取得する。次に現在参照中の要素のパス名称IDの取得を行なう（ステップ403）。図6は図3のような木構造を持つ構造化文書を登録した時に、最終的に作成されるパス名称インデックスの例である。パス名称インデックスは、登録対象文書のパス名称に一意なID（パス名称ID）を割当てたものである。また各パス名称IDは最下層のタグ名の名称IDの情報を持つ。図3の301の要素のパス名称は「/論文/本文/章/節/段落」であり、このパス名称に割当てられたパス名称IDは、図6の例では601に示される値（N11）である。ステップ403では、このパス名称インデックスに現在参照中の要素に該当するパス名称のノードが存在する場合はそのパス名称IDを取得し、存在しない場合にはそのパス名称のノードとそのパス名称IDを新たに作成し、パス名称インデックス格納手段116に格納するとともに、そのパス名称IDを取得する。なお、ここでパス名称を表現する際に、各階層の区切り文字として“/”（スラッシュ）を用いたが、これはタグ名に用いられない文字である限りどのようなものでも構わない。次に現在参照中の要素のパス階層IDの取得を行なう（ステップ404）。図7は図3のような木構造を持つ構造化文書を登録した時に、最終的に作成されるパス階層インデックスの例である。パス階層インデックスは、登録対象文書のパス階層に一意なID（パス階層ID）を割当てたものである。図3の301の要素のパス階層は「/1/1/1/1/2」であり、このパス階層に割当てられたパス階層IDは図7の例では701に示される値（L5）である。ステップ404では、このパス階層インデックスに現在参照中の要素に該当するパス階層のノードが存在する場合はそのパス階層IDを取得し、存在しない場合にはそのパス階層のノードとそのパス階層IDを新たに作成し、パス階層インデックス格納手段117に格納するとともに、そのパス階層IDを取得する。なお、ここでパス階層を表現する際に、パス名称と同様に各階層の区切り文字として“/”（スラッシュ）を用いたが、これは出現順序を表す数字に用いられない文字である限りどのようなものでも構わない。次に現在参照中の要素が実体を持つかどうかチェックし（ステップ405）、実体を持たない場合はステップ408へ進む。実体を持つ場合、ステップ406へ進む、この要素に検索単位識別子を割当てる。次にステップ407で要素管理テーブルに現在参照中の要

素のレコードを追加する。図8は要素管理テーブルの例であり、801は図3の301の要素に関するレコードに該当する。実施の形態1における要素管理テーブルは、検索単位識別子をキーとして文書番号、パス名称ID、パス階層ID、名称IDを管理する。次にステップ408で登録対象文書の全要素についてステップ402から407の処理を終了したか調べ、まだ未処理の要素が存在したらステップ402以降の処理を繰り返す。

【0062】次に文字列索引作成手段109では、各要素ごとに要素内容の検索用文字列索引を作成する。文字列索引作成手段109の処理の流れを図9を用いて説明する。

【0063】まず構造解析済みデータ格納手段114から登録対象文書の構造解析済みデータを読み込む(ステップ901)。次に現在参照中の要素が実体を持つかどうかチェックし(ステップ902)、実体を持たない場合はステップ807へ進む。実体を持つ場合、ステップ903へ進み、構造情報作成手段108の処理ステップ406でこの要素に割当てた検索単位識別子を取得する。次に該要素内容の文字列についてあらかじめ定めた文字数の文字連鎖を取り出す(ステップ904)。

【0064】この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号(以下、文字位置番号と呼ぶ)の情報を文字列索引に追加する(ステップ905)。ステップ904、905の処理を該要素の全文字列について繰り返す(ステップ906)。次にステップ907で登録対象文書の全要素についてステップ902から906の処理を終了したか調べ、まだ未処理の要素が存在したらステップ902以降の処理を繰り返す。

【0065】全要素についてステップ902から906の処理を終了したら、最後にここで作成した文字列索引を文字列索引格納手段119に追加する(ステップ908)。

【0066】図10は文字列索引作成手段109によって図2の構造化文書のうち3行目の「<；タイトル> 構造化文書管理 <；／タイトル>」という要素について作成した文字列索引の例の一部を示した図である。図10の1001は「検索単位識別子が“1”の要素の文字列中に“構造”という文字連鎖が先頭から“1”文字目の位置から存在する」ということを表している。なお、図10は文字列索引の一部しか示していないが、実際は登録対象文書の全要素の全文字列について文字列索引が作成される。

【0067】なお、この例では2文字ずつ文字連鎖を取り出してそれぞれに文字列索引を作成しているが、この文字連鎖は2文字ずつでなくても構わない。また、以上の登録処理を登録対象文書が入力されるごとに繰り返すことにより、構造情報と文字列索引が追加されてゆく。

【0068】なお、図5他において名称ID、パス名称

IDおよびパス階層IDは“T9”や“N11”や“L5”といった文字で表現しているが、これらはそれぞれ、名称(タグ名)を一意に特定するID、パス名称を一意に特定するID、パス階層を一意に特定するIDであればどのようなものでも構わない。次に本実施の形態における文書検索の処理の流れを具体例を示して説明する。

【0069】なお、以下に示す本実施の形態における文書検索処理の説明においては、名称IDテーブル、パス名称インデックス、パス階層インデックス、要素管理テーブルには、それぞれ図5、図6、図7、図8のようなデータが格納されているものとして説明を行なう。

【0070】まず検索条件入力手段103を通して、端末101から「パス名称が“／論文／書誌／タイトル”である要素に、“構造化”という文字列が含まれる文書」という条件が与えられたとする。

【0071】図11は検索条件解析手段110の処理の流れを示した図である。ここでの例は、検索条件の構造指定としてパス名称のみ指定されているので、図11のCase3に該当する。Case3ではステップ1102で、パス名称インデックス格納手段116に格納されているパス名称インデックスを参照し、検索条件のパス名称をパス名称IDに変換する。パス名称インデックスが図6の場合、検索条件のパス名称“／論文／書誌／タイトル”は、パス名称ID“N2”に変換される。

【0072】次に文字列索引検索手段111で、検索条件の文字列について検索処理を行なう。図12は文字列索引検索手段111での処理を図に示したものである。ここでの例では検索条件の文字列は“構造化”であり、これは2文字ずつの文字連鎖として“構造”と“造化”が取り出せる。ここで取り出す文字連鎖の文字数は、文字列索引作成手段109で作成する文字連鎖の文字数と同一とする。この2つの文字連鎖について図12の1210に示すような文字列索引が作成されているとして、この中から検索単位識別子が同一で、かつ“構造”の文字連鎖から“造化”の文字連鎖に対して文字位置番号が連続しているものを文字列索引検索手段111の結果として抽出する。図12の例では検索単位識別子が同一なものとして1221、1222、1223を取り出すことが出来る。更にその中で文字位置番号が連続しているのは1221と1223であり、これらの検索単位識別子を抽出する。

【0073】次に構造照合手段112で、文字列索引検索手段111で得られた検索単位識別子群の中から、検索条件の構造指定を満たす最終的な検索結果を求める。図13は、構造照合手段112の処理の流れを示した図である。図13におけるCase1からCase4は、図11の検索条件の構造指定パターンCase1からCase4と同様である。ここでの例ではCase3(パス名称のみ指定)であるので、ステップ1303でパス

名称の照合を行なう。図14はこの例における構造照合処理の詳細を示す図である。まず文字列索引検索手段111で得られた検索単位識別子(1401)をキーとして要素管理テーブルを参照する。そこで該検索単位識別子のパス名称IDが、検索条件解析手段110で求めた検索条件のパス名称ID(この例では“N2”)と一致するものだけを最終的な検索結果とする。

【0074】なお、本実施の形態では検索条件の構造指定として、タグ名を指定した検索(Case1)、タグ名とその出現順序を指定した検索(Case2)、パス名称とパス階層を指定した検索(Case4)にも対応可能である。以下でそれぞれCaseでの処理について簡潔に説明する。

【0075】タグ名を指定した検索(Case1)の場合、まず図11より検索条件解析手段110にて、検索条件のタグ名を名称IDに変換する(ステップ1101)。

【0076】次にCase3と同様に、文字列索引検索手段111にて検索条件の文字列について検索処理を行ない、該当する検索単位識別子群を求める。最後に図13より構造照合手段112にて、文字列索引検索手段111で求めた検索単位識別子群のうち、名称IDがステップ1101で求めた名称IDと一致するものだけを、要素管理テーブルを元に抽出し(ステップ1301)、最終的な検索結果とする。

【0077】タグ名とその出現順序を指定した検索(Case2)の場合、Case1と同様な処理を行なった後、最後に出現順序照合処理(図13のステップ1302)を行なう。ステップ1302では、該検索単位識別子のパス階層IDをキーとしてパス階層インデックスを参照し、末端階層の出現順序が検索条件の出現順序と一致するものだけを抽出し、最終的な検索結果とする。

【0078】パス名称とパス階層を指定した検索(Case4)の場合、検索条件解析手段110でCase3と同様にステップ1102の処理を行なった後、検索条件のパス階層をパス階層インデックスを用いてパス階層IDへの変換を行なう(ステップ1103)。次にCase3と同様に、文字列索引検索手段111にて検索条件の文字列について検索処理を行ない、該当する検索単位識別子群を求める。

【0079】最後に構造照合手段112にて、Case3と同様にパス名称ID照合処理(ステップ1303)を行なった後、パス階層ID照合処理(ステップ1304)を行なう。ステップ1304では、該検索単位識別子のパス階層IDがステップ1103で変換したパス階層IDと一致するものだけを抽出し、最終的な検索結果とする。

【0080】最後に検索結果作成・表示処理について説明する。結果作成手段113は検索結果として得られた文書の書誌情報(タイトル、著者、日付など)を結果一

覧表示用のデータとして、一覧データ格納手段121に格納する。このデータを結果表示手段104を通して端末101に表示する。次に端末101から実体表示要求としてこの検索結果一覧の中からどれか1つの文書が選択されると、結果作成手段111が実体データ格納手段115から指定された文書の実体を取得し、結果表示手段104を通して端末101に表示する。なお、構造解析手段107によって要素に分割された単位で、登録対象文書を実体データ格納手段120に登録しておくことにより、検索結果作成・表示処理において要素毎の結果一覧の作成、および要素毎の実体取得も可能である。

【0081】以上のように本実施の形態では、構造化文書の論理構造情報を要素管理テーブル格納手段115、パス名称インデックス格納手段116、パス階層インデックス格納手段117、名称IDテーブル格納手段118の4つに分けて格納し、文字列索引内部にこれら論理構造に関する情報を含めないことにより、文字列索引のサイズ縮小を可能とする。更に文書の特定の要素内容の追加、変更、削除を行なう際に、追加、変更、削除により論理構造の変化の発生した検索単位識別子のレコードについて、要素管理テーブルの変更処理を行なうだけで済むため、文字列索引内部に論理構造に関する情報を含める方法と比較して、処理量の大幅な軽減が可能となる。(文字列索引内部に論理構造に関する情報を含める方法の場合、追加、変更、削除により、論理構造の変化が発生した要素に関する全文字連鎖の文字列索引に対して修正処理が発生するため。)具体例を以下に示す。図15は図3の構造をした文書の第1章第1節と第1章第2節の間に1501に示すノード群を追加した例である。この場合、1502のノードは第1章第2節から第1章第3節へと変更しなくてはならない。この時本実施の形態の方法では、既登録のデータに関しては、要素管理テーブルにおける検索単位識別子10、および11のレコードのパス名称IDとパス階層IDを変更するだけで済む。一方、文字列索引内部に論理構造に関する情報を含める方法の場合、検索単位識別子10および11の要素の全文字連鎖の文字列索引に対して論理構造情報の変更を行わなくてはならない。(仮に、検索単位識別子10の要素の内容が100文字であったとすると、2文字連鎖で索引を作成している場合、99個の文字連鎖の文字列索引に対して変更が必要となる)。

【0082】また、本実施の形態では要素の論理構造位置を特定するためのIDをパス名称IDとパス階層IDの2つに分けているため、論理構造が複雑かつ膨大になった場合でも、公知例のように1種類のID(文脈識別子)で論理構造位置を特定する方法と比較して、IDの総数を少なく押さえることが可能となる。

【0083】なお、本実施の形態では1文書の構造化文書の登録、検索について説明したが、複数文書の場合でも同様の処理で実現が可能である。また本実施の形態で

は、一種類のDTDにおけるパス名称IDの作成方法について説明したが、本システムに複数の異なるDTDの文書の登録要求が起こった場合においても、各ノードに個別なパス名称IDを割当てることにより、論理構造を指定した検索が実現可能である。また、要素管理テーブル、パス名称インデックス、パス階層インデックス、名称IDテーブルを一次記憶上に持つことにより、構造照合手段112の高速化が可能である。

【0084】また本実施の形態は、構造化文書の管理を目的とする装置について説明を行ったが、必ずしも構造化文書に限らず、木構造で表現可能なデータを管理するために上述のパス名称インデックス及びパス階層インデックスを利用して実体要素（データの实体）を管理することも可能である。

【0085】さらに実施の形態1は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【0086】（実施の形態2）以下、本発明の実施の形態2について説明する。図16は実施の形態2における構造化文書管理装置の構成図である。実施の形態1の構成図である図1と異なるのは、データ格納部106にパス名称ID照合テーブル格納手段1601、パス階層ID照合テーブル格納手段1602を新たに備えているところである。またそれに伴い、検索条件解析手段110、および構造照合手段112の処理が実施の形態1とは異なる。

【0087】パス名称ID照合テーブル格納手段1601は、各パス名称IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納される。

【0088】パス階層ID照合テーブル格納手段1602は、各パス階層IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納される。

【0089】実施の形態2における目的は、実施の形態1における検索条件の構造指定パターンCase1からCase4以外の構造指定に対応することである。Case1からCase4はタグ名やパス名称などで指定された末端要素そのものに対して検索を行なうものである。実施の形態2で実現する検索は、実体を持たない中間ノード以下を指定した検索である。例えば、「“章”以下に“管理”という文字列を含む文書を検索する」といった検索条件に対応することを目的とする。

【0090】実施の形態2における登録処理は、実施の形態1と同様であるため説明を省略する。

【0091】次に実施の形態2における検索処理の流れを具体例を示して説明する。なお、以下に示す本実施の形態における文書検索処理の説明においては、名称IDテーブル、パス名称インデックス、パス階層インデックス、要素管理テーブルには、それぞれ図5、図6、図

7、図8のようなデータが格納されているものとして説明を行なう。

【0092】まず、検索条件入力手段103を通して、端末101から「パス名称が“／論文／本文／章”である中間ノード以下である要素に、“管理”という文字列が含まれる文書」という条件が与えられたとする。

【0093】図17は実施の形態2における検索条件解析手段110の処理の流れを示した図である。ここでの例では検索条件の構造指定としてパス名称以下が指定されているので、図17のCase7に該当する。Case7ではステップ1102で、実施の形態1と同様に検索条件のパス名称をパス名称IDに変換する。パス名称インデックスが図6の場合、検索条件のパス名称“／論文／本文／章”はパス名称ID“N6”に変換される。次にステップ1701でパス名称ID照合テーブルを作成する。図18はここでの検索条件の例におけるパス名称ID照合テーブルの内容を示す図である。このパス名称ID照合テーブルは、検索要求ごとに作成し、パス名称インデックスの全パス名称IDについて、検索条件で指定された範囲内のパス名称IDと範囲外のパス名称IDを即座に判断するために作成する。この例の場合、図6のパス名称インデックスよりパス名称ID“N6”以下にあるパス名称ID“N7、N8、N9、N10、N11”が範囲内で、それ以外は範囲外となる。

【0094】次に文字列索引検索手段111で、検索条件の文字列について検索処理を行なう。処理手順は実施の形態1と同様であるため省略するが、ここでの例である“管理”という文字列で検索した結果として、検索単位識別子“1”と“9”が得られたものとして、説明を続ける。

【0095】次に構造照合手段112で、文字列索引検索手段111で得られた検索単位識別子群の中から、検索条件の構造指定を満たす最終的な検索結果を求める。図19は実施の形態2における構造照合手段112の処理の流れを示した図である。

【0096】図19におけるCase5からCase8というのは、図17の検索条件の構造指定パターンCase5からCase8と同様である。ここでの例では、Case7（パス名称以下を指定）であるので、ステップ1303のパス名称ID照合処理を行なう。ただし、Case7におけるパス名称ID照合処理は、パス名称ID照合テーブルを用いて照合を行なう。図20はこの例における構造照合処理の詳細を示す図である。まず文字列索引検索手段111で得られた検索単位識別子群（2001）をキーとして要素管理テーブルを参照する。そこで該検索単位識別子のパス名称IDからパス名称ID照合テーブルを参照し、照合フラグが“1”（範囲内）であるものだけを最終的な検索結果とする。

【0097】なお、本実施の形態では、検索条件の構造指定として、タグ名で指定された中間ノード以下に対す

る検索（Case 5）、タグ名とその出現順序で指定された中間ノード以下に対する検索（Case 6）、パス名称とパス階層で指定された中間ノード以下に対する検索（Case 8）にも対応可能である。以下でそれぞれCaseでの処理について簡潔に説明する。

【0098】タグ名で指定された中間ノード以下に対する検索（Case 5）の場合、検索条件解析手段110と文字列索引検索手段111における処理は、実施の形態1のCase 1と同様であるため省略する。最後に図19より構造照合手段112にて構造指定のチェックを行なう。ここでステップ1901のパス名称ID作成・更新・照合処理について説明する。図21はパス名称ID作成・更新・照合処理の流れを示したフローチャートであり、このフローチャートに沿って説明する。

【0099】まずパス名称ID照合テーブルの照合フラグを“0”（未定）で初期化しておく（ステップ3101）。次に文字列索引検索手段111で求めた検索単位識別子群それぞれについて以下の処理を繰り返す。まず検索単位識別子を取得し（ステップ3102）、該検索単位識別子のパス名称ID（要素管理テーブルより取得）の照合フラグを参照（ステップ3103）し、該照合フラグが“1”（範囲内）であれば（ステップ3104）、該検索単位識別子を最終的な検索結果に含める（ステップ3105）。照合フラグが“2”（範囲外）であれば（ステップ3106）、該検索単位識別子は最終的な検索範囲に含めない（ステップ3107）。照合フラグが“0”（未定）であったら、該検索単位識別子のパス名称IDをキーとしてパス名称インデックスを参照し（ステップ3108）、検索条件解析手段110のステップ1101で求めた名称IDと一致するか、もしくは、たどったノードのパス名称IDの照合フラグが“1”（範囲内）の場合（ステップ3109）、該検索単位識別子のパス名称IDと、そこまでたどったパス名称ID全てに対して、パス名称ID照合テーブルの照合フラグを1に設定し（ステップ3110）、該検索単位識別子を最終的な検索結果に含める。

【0100】逆に、たどったノードのパス名称IDの照合フラグが“2”（範囲外）の場合（ステップ3111）、該検索単位識別子のパス名称IDと、そこまでたどったパス名称ID全てに対して、パス名称ID照合テーブルの照合フラグを“2”（範囲外）に設定し（ステップ3112）、該検索単位識別子を最終的な検索結果に含めない。

【0101】さらに、たどったノードのパス名称IDの照合フラグが“0”（未定）の場合は、1階層登り（ステップ3113）、ルートノードであるか否かを判定し（ステップ3114）し、ルートノードでなければ、再びステップ3108に戻る。ルートノードである場合は、該検索単位識別子のパス名称IDと、それまでたどったパス名称ID全ての照合フラグを2“範囲外”に設

定する（ステップ3112）。

【0102】次の該当検索単位識別子が存在するか否かをチェックし（ステップ3115）、存在する場合は、ステップ3102へ戻る。存在しない場合は、本処理を終了する。

【0103】このように徐々に各パス名称IDが検索条件の範囲内にあるかどうかのパス名称ID照合テーブルが学習されていくため、別の検索単位識別子に対してパス名称IDの照合を行なう際に、すでに範囲内であると判明している（照合フラグが“1”である）パス名称IDであった場合、該検索単位識別子を即座に最終的な検索結果に含ませることが可能となる。

【0104】なお上記ステップ3101からステップ3115までの処理については、汎用計算機に上記ステップの処理を実現するプログラムをインストールすることにより実現することが可能である。

【0105】また上記実施の形態では、構造化文書において中間ノード以下を検索範囲に指定した場合に、検索範囲に含まれるノードを決定する例を示したが、構造化文書に限らず、その他木構造で表現できるデータについても同様に適用することが可能である。

【0106】タグ名とその出現順序で指定された中間ノード以下に対する検索（Case 6）の場合、検索条件解析手段110、文字列索引検索手段111、および構造照合手段112のステップ1901まではCase 5と同様の処理を行なう。次にステップ1901でパス名称IDが範囲内にあった場合に限り、ステップ1902のパス階層ID作成・更新・照合処理を行なう。図22はパス階層ID照合テーブルの例である。ステップ1902ではステップ1901のパス名称IDに関する処理と同様に、パス階層IDについて構造指定の範囲にあるかどうか学習していき、照合フラグが“1”のパス階層IDを持つ検索単位識別子を最終的な検索結果とする。

【0107】パス名称とパス階層で指定された中間ノード以下に対する検索（Case 8）の場合、検索条件解析手段110では、Case 7と同様な処理を行なったあとに、ステップ1702にてパス階層ID照合テーブルを作成する。このパス階層ID照合テーブルは、パス階層インデックスにおいて、ステップ1103で求めたパス階層IDにあたるノードとそれ以下全てのノードのパス階層IDに対する照合フラグを“1”（範囲内）に、それ以外を“2”（範囲外）に設定する。文字列索引検索手段111での処理はCase 7と同様であるため説明を省略する。

【0108】次に構造照合手段112において、Case 7と同様な処理を行なった後、ステップ1702にて作成したパス階層ID照合テーブルを用いて、該検索単位識別子のパス階層IDの照合処理を行なう。ここでパス階層ID照合テーブルの照合フラグが“1”であるパス階層IDを持つ検索単位識別子のみ、最終的な検索結

果とする。

【0109】実施の形態2における検索結果作成・表示処理は実施の形態1と同様であるため、説明を省略する。

【0110】以上のように本実施の形態では、中間ノードを以下を指定した検索の際に、各パス名称IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス名称ID照合テーブルや、各パス階層IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス階層ID照合テーブルを作成し、構造照合処理を行なうことにより、中間ノード以下を指定した高速な検索を実現する。

【0111】なお、図16に示す実施の形態2の構成においても、パス名称ID照合テーブル格納手段1601、およびパス階層ID照合テーブル格納手段1602を使用しないことにより、実施の形態1における検索条件の構造指定Case1からCase4にも、対応可能である。また本実施の形態の説明において、パス名称ID照合テーブル、およびパス階層ID照合テーブルの照合フラグの値を、範囲内の場合“1”、範囲外の場合“2”、未定の場合“0”としていたが、この照合フラグの値は範囲内、範囲外、未定の状態を判別可能な値であればどのような値を割当てても構わない。

【0112】さらに実施の形態2は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【0113】（実施の形態3）以下、本発明の実施の形態3について説明する。実施の形態3における構造化文書管理装置の構成図は実施の形態1における図1、もしくは実施の形態2における図16と同様である。ただし、文字列索引作成手段109における文字列索引の作成方法が実施の形態1および実施の形態2とは若干異なり、それに伴い文字列索引検索手段111と構造照合手段112における処理が実施の形態1および実施の形態2とは異なる。

【0114】ここで実施の形態3における登録処理の流れについて説明する。まず構造化文書入力手段102、構造解析手段107、および構造情報作成手段108の処理は、実施の形態1および実施の形態2と同様であるため説明を省略する。

【0115】図23は実施の形態3における文字列索引作成手段109の処理の流れである。ステップ901からステップ903までは実施の形態1および実施の形態2と同様であるため説明を省略する。次に該要素がMixed Contentを含むかどうか調べ（ステップ2201）、含む場合はこのMixed Contentに割当てられている検索単位識別子を取得する（ステップ2202）。この「Mixed Content」

とは、要素実体の内部で、該要素の子要素として存在する、要素実体のことである。例えば、図24の2310のように、「段落」を表す要素の中で、更に「キーワード」タグに囲まれた要素がMixed Contentである。他の例としては、「強調」や「斜体」などがあり、検索する際には、これら「段落」と「キーワード」の要素にまたがった文字列でも検索してヒットすることが望まれる。そのためステップ2203で文字連鎖を取り出す際に、Mixed Contentにまたがる文字連鎖も抽出し、Mixed Contentにまたがる文字連鎖の場合には、ステップ2204で文字列索引に、文字連鎖1文字目の検索単位識別子と文字連鎖2文字目の検索単位識別子と文字位置番号を格納する（以下、このようなMixed Contentにまたがる文字連鎖の文字列索引を、拡張文字列索引と呼ぶ）。この場合の文字位置番号は、該文字連鎖先頭文字がMixed Contentの外側の要素の中で何番目の文字かを表す番号とする。ステップ906から908までの処理は、実施の形態1および実施の形態2と同様であるため説明を省略する。

【0116】次にMixed Contentを含む要素の文字列索引の作成例について、図24を用いて説明する。図24の2310に示すように、「段落」の中に「キーワード」タグで囲まれたMixed Contentを含み、「キーワード」タグの要素の検索単位識別子は“101”、「段落」タグの要素の検索単位識別子は“102”が割当てられているものとして説明する。この例の場合に作成される文字列索引を図示したものが2320である。この例の場合、“を”（2321）と“索す”（2323）の文字連鎖がMixed Contentにまたがっており、文字連鎖1文字目と文字連鎖2文字目の、2個の検索単位識別子が文字列索引に格納される。なお、図24の2320は文字列索引の一部しか示されていないが、実際は登録対象文書の全要素の全文字列について文字列索引が作成される。

【0117】次に実施の形態3における文書検索の処理の流れについて説明する。まず検索条件入力手段103、検索条件解析手段110における処理は実施の形態1および実施の形態2と同様であるため説明を省略する。次に文字列索引検索手段111における処理についてだが、基本的には実施の形態1および実施の形態2と同様である。ただし実施の形態3では、文字列索引作成手段109において、Mixed Contentにまたがる文字連鎖の場合、文字連鎖1文字目と文字連鎖2文字目の、2個の検索単位識別子含む拡張文字列索引を作成しているため、この拡張文字列索引が絡む場合の検索処理が新たに必要となる。以下、その具体例について図24を用いて説明する。検索文字列が“検索する”である場合、2310の要素に該当する文字連鎖の文字列索引として2322、2323、2324が得られる。

ここで2322の検索単位識別子と、拡張文字列索引である2323の文字連鎖1文字目検索単位識別子が“101”で一致する。更に文字位置番号が“4”と“5”で連続している。また、拡張文字列索引2323の文字連鎖2文字目検索単位識別子と2324の検索単位識別子が“102”で一致し、更に文字位置番号が“5”と“6”で連続している。このような場合に文字連鎖2333から2324にかけて文字列検索にヒットしたことになる。その際、文字列検索結果の検索単位識別子として、検索文字列の先頭文字および末尾文字に該当するの検索単位識別子のセットを返す。ここでの例の場合、先頭文字検索単位識別子“101”、末尾文字検索単位識別子“102”のセットを返す。次に構造照合手段の処理についてだが、基本的には実施の形態1および実施の形態2と同様である。ただし実施の形態3では、文字列索引検索手段111から得られる文字列検索結果群の中に、先頭文字検索単位識別子と末尾文字検索単位識別子のセットが含まれる場合があり、この場合の構造照合処理が新たに必要となる。

【0118】上記実施の形態3における文字列索引検索手段111の説明で用いた例では、文字列検索処理結果として、先頭文字検索単位識別子“101”、末尾文字検索単位識別子“102”のセットを返した。この場合、検索単位識別子“101”および“102”の両方に対して、実施の形態1および実施の形態2と同様な構造照合処理を行ない、両検索単位識別子とも検索条件の構造指定に当てはまる場合のみ、最終的な検索結果とする。

【0119】実施の形態3における検索結果作成・表示処理は実施の形態1および実施の形態2と同様であるため、説明を省略する。

【0120】以上のように本実施の形態では、登録対象構造化文書中にMixed Contentを含む場合に、Mixed Contentにまたがる文字連鎖に対しても文字列索引（文字連鎖1文字目と文字連鎖2文字目の、2個の検索単位識別子を記憶する拡張文字列索引）を作成することによって、Mixed Contentにまたがる文字列を検索対象とすることが可能となる。また、Mixed Contentである要素（上記説明では「キーワード」要素）を指定した検索も可能となる。

【0121】なお、実施の形態3の説明においては、2文字ずつ文字連鎖を取り出してそれぞれに文字列索引を作成しているが、この文字連鎖は2文字ずつでなくても構わない。この場合、実施の形態3における「文字連鎖1文字目検索単位識別子」を「文字連鎖先頭文字の検索単位識別子」に、「文字連鎖2文字目検索単位識別子」を「文字連鎖末尾文字の検索単位識別子」に置き換えることにより、同様の効果が実現可能である。

【0122】さらに実施の形態3は、装置として実現す

る例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【0123】（実施の形態4）以下、本発明の実施の形態4について説明する。図25は実施の形態4における構造化文書管理装置の構成図である。実施の形態1の構成図である図1と異なるのは、検索エンジン105に数値型索引作成手段2401と数値型索引検索手段2402を、データ格納部106に数値型設定格納手段2403と数値型索引格納手段2404を新たに備えているところである。

【0124】数値型索引作成手段2401は、あらかじめ設定されたタグ名の要素内容に対する数値範囲検索用の索引を作成する。

【0125】数値型索引検索手段2402は、数値型索引作成手段2401で作成された数値型索引を用いて数値範囲の検索処理を行なう。

【0126】数値型設定格納手段2403は、あらかじめ数値型索引を作成するように定められた要素のタグ名の集合が格納されている。

【0127】数値型索引格納手段2404は、数値型索引作成手段2401で作成された数値型索引を格納する。

【0128】ここで、実施の形態4における登録処理の流れについて具体例を用いて説明する。まず実施の形態4においては、本システムに文書を登録する前に、数値型設定格納手段2403にあらかじめ数値索引を作成する要素のタグ名として“価格”というタグ名が設定されているものとする。この時、図26のような文書を登録する場合について説明する。構造化文書入力手段102、構造解析手段107、構造情報作成手段108、および文字列索引作成手段109の処理は、実施の形態1および実施の形態2と同様であるため説明を省略する。

【0129】図27は実施の形態4における数値型索引作成手段2401の処理の流れである。まずステップ2601で登録文書の構造解析済みデータを読み込む。次に現在参照中の要素が数値型設定格納手段2403で数値型索引を作成するよう設定された要素かどうか調べ（ステップ2602）、設定されていない要素であったらステップ2606へ進む。設定されていた要素であったら、構造解析手段107のステップ406にて該要素に割当てられた検索単位識別子を取得する。次にステップ2604で該要素の実体（文字列）を数値データに変換する。その際、文字列が数字だけでなく単位などの文字データを含んでいる場合、数字部分の文字列だけ取り出し、数値データに変換する。そして数値型索引に該要素の検索単位識別子と数値データのレコードを追加する。この際、数値型索引は数値型設定格納手段2403で設定された要素のタグ名の名称IDごとに作成する

(ステップ2605)。次にステップ2606で登録対象文書の全要素についてステップ2602から2605の処理を終了したか調べ、まだ未処理の要素が存在したらステップ2602以降の処理を繰り返す。全要素についてステップ2602から2605の処理を終了したら、最後にここで作成した数値型索引を数値型索引格納手段2404に追加する(ステップ2607)。

【0130】ここでの例の場合、数値型索引を作成する要素は図26の2501に示す要素である。該要素の検索単位識別子が“201”であるとした場合に作成される数値型索引は図28の2710のようになる。なお、図28では数値データをLong型整数として格納しているが、Double型浮動小数点数などで格納することも可能である。ただし、名称ID単位で作成される数値型索引ごとに型を統一する必要がある。

【0131】次に実施の形態4における文書検索の処理について説明する。実施の形態4では、数値型設定格納手段2403で設定されたタグ名の要素に対して数値型索引を作成しているため、実施の形態1および実施の形態2で説明した構造を指定した文字列の検索のほかに、数値範囲の検索が可能となる。

【0132】例として、まず検索条件入力手段103を通して、端末101から「タグ名が“価格”である要素の内容が“1500円～1700円”である文書」という条件が与えられたとする。この時検索条件解析手段110の処理は実施の形態1のCase1と同様であるため説明を省略する。

【0133】次に検索条件が数値範囲を指定した検索なので、文字列索引検索手段111ではなく、数値型索引検索手段2402の処理を行なう。ここでの例の場合、“価格”タグの名称IDについて作成された数値型索引に図28の2720のようなデータが格納されているとすると、1500以上、1700以下の数値データを持つものとして2721(検索単位識別子:54)、2722(検索単位識別子:201)、2723(検索単位識別子:545)の3つを抽出する。

【0134】次に構造照合手段112にて、数値型索引検索手段2402の処理で抽出した検索単位識別子について、検索条件の構造指定チェックを行なう。ここでの例における構造照合手段112の処理は実施の形態1と同様であるため説明を省略する。なお、実施の形態4では数値範囲検索における構造指定として、上記実施の形態1におけるCase1のみでなく、Case2、Case3、Case4に対応可能である。それぞれのCaseにおける検索条件解析手段110および構造照合手段112における処理は、実施の形態1と同様であるため説明を省略する。

【0135】実施の形態4における検索結果作成・表示処理は実施の形態1と同様であるため、説明を省略する。

【0136】以上のように本実施の形態では、あらかじめ数値型設定格納手段2403で設定されたタグ名の要素に対して数値型索引作成手段2401にて数値型索引を作成することにより、要素内容を数値データとして扱った数値範囲の検索が可能となる。

【0137】なお実施の形態4における数値型索引は、図28の2720のような構造であるとして説明したが、この数値型索引は指定された数値範囲に該当する検索単位識別子を抽出できるものであればどのような構造でも構わない。また、実施の形態4において、文字列索引作成手段109での処理の後に、数値型索引作成手段2401を行なうものとして説明したが、文字列索引作成手段109の処理手順である図4のステップ405にて、要素実体に会った場合に、ステップ406と407の処理と平行して、数値型索引作成手段2401の処理手順である図27のステップ2602からステップ2605の処理を行なうことも可能である。

【0138】さらに実施の形態4は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【0139】(実施の形態5)以下、本発明の実施の形態5について説明する。図29は実施の形態5における構造化文書管理装置の構成図である。

【0140】本実施の形態は、ネットワーク上に構造化文書管理装置の各機能が分散していることを特徴とするものである。

【0141】構造化文書登録部3001は、構造化文書を読み込み、解析し、構造化文書の本構造を生成する機能を有している。文字列索引作成部3002は、構造化文書登録部3001で解析された構造化文書について、検索用索引を生成する機能を有している。文字列検索部3003は、検索条件を読み込み、検索条件に該当する文字列を有している要素実体を検索する機能を有している。結果表示部3004は、前記文字列検索部3003で得られた検索結果を端末101に表示する機能を有している。なお、端末101及びデータ格納部106は実施の形態1で記載した機能と同一の機能を有しており、データ格納部106は上記各機能ブロックが作成した解析済構造化文書、文字列索引、検索結果等をネットワーク経由で受け取り記憶する。端末101は、使用者の指定した検索条件をネットワーク経由で文字列検索部3003に送る。また、結果表示部3004に記憶されている検索結果をネットワーク経由で受け取り、表示する機能を有している。以下、各機能ブロック毎に説明する。

【0142】構造化文書登録部3001は、構造化文書入力手段102と構造解析手段107と構造情報作成手段108より構成されており、これら3つの手段は、実施の形態1で記載している機能と同一の機能を有してい

る。ただし、構造情報作成手段108で作成される要素管理テーブルは実施の形態1に記載した図8の形式の他、図31または32のように検索単位識別子とパス名称ID及びパス階層IDとの対応関係を示した形式、または検索単位識別子と名称IDとの対応関係を示した形式でも構わない。

【0143】なお、上記構造化文書登録部3001の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、構造化文書登録部3001と同一の機能を実現できる。

【0144】また上記構造化文書登録部3001は、それ自体で装置としての機能も果たすことが可能である。

【0145】文字列索引作成部3002は、文字列索引作成手段109と、数値型索引作成手段2401から構成されている。文字列索引作成手段109は実施の形態1に記載した機能と同一の機能を有している。数値型索引作成手段2401は実施の形態4に記載した機能と同一の機能を有している。ただし、数値型索引作成手段2401は、検索条件として特定の数値範囲に該当する文字列を検索する場合に必要となる構成要素であり、検索条件に数値範囲が含まれない場合は、数値型索引作成手段2401は不要である。

【0146】なお、上記文字列索引作成部3002の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、文字列索引作成部3002と同一の機能を実現できる。

【0147】また文字列索引作成部3002は、それ自体で装置としての機能も果たすことが可能である。

【0148】文字列検索部3003は、検索条件入力手段103と、検索条件解析手段110と、文字列索引検索手段111と、数値型索引検索手段2402と、構造照合手段112から構成されている。検索条件入力手段103、検索条件解析手段110と、文字列索引検索手段111と、構造照合手段112は、実施の形態1に記載の機能と同一の機能を有する。ただし、構造情報作成手段108で作成される要素管理テーブルが図31の形式の場合は、検索条件としてタグ名を指定することはできず、パス名称またはパス階層を指定することができる。一方、要素管理テーブルが図32の形式の場合は、検索条件としてタグ名のみを指定することができる。

【0149】なお、上記文字列検索部3003の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、文字列索引部3403と同一の機能を実現できる。

【0150】また文字列検索部3003は、それ自体で装置としての機能も果たすことが可能である。

【0151】数値型索引検索手段2402は実施の形態

4に記載の機能と同一の機能を有する。ただし、数値型索引検索手段2402は、検索条件として特定の数値範囲に該当する文字列を検索する場合に必要となる構成要素であり、検索条件に数値範囲が含まれない場合は、数値型索引検索手段2402は不要である。

【0152】なお、上記数値型索引検索手段2402の機能と同一の機能はプログラム形式で実行可能であり、このプログラムを記録した可搬型媒体を用いて汎用計算機にインストールすることにより、数値型索引検索手段2402と同一の機能を実現できる。

【0153】図30は、文字列検索部3003の処理の流れを示したフローチャートである。

【0154】まず、使用者の指定した検索条件を読み込み（ステップ3005）、次に、読み込んだ検索条件に該当する名称ID又は、パス名称ID又は、パス階層IDのいずれかのID（以下ID1）に変換する（ステップ3006）。なお、前記3つのIDのうち、いずれのIDに変換されるかは図11に示すように使用者の検索条件に依存する。また、どのような検索条件が可能であるかは図8、図31、図32に示した要素管理テーブルの形式に制約される。次に、前記検索条件に該当する文字列を有するすべての検索単位識別子（以下、ID2）を特定する（ステップ3007）し、前記ID2に基づいて要素管理テーブルを参照し、対応する名称ID又は、パス名称ID又は、パス階層IDのいずれかのID（以下、ID3）を特定し（ステップ3008）、最後に、前記ID1とID3が一致する検索単位識別子を特定する（ステップ3009）。

【0155】結果表示部3004は、結果作成手段113と結果表示手段104から構成されている。結果作成手段113と結果表示手段104は、実施の形態1に記載の機能と同一である。

【0156】さらに実施の形態5は、装置として実現する例を示したが、その他に汎用計算機に本実施の形態に開示した構造化文書管理装置として機能するプログラムをインストールすることによっても実現することが可能である。

【0157】

【発明の効果】以上のように、本発明によれば構造化文書の様々な論理構造を指定した検索が可能な構造化文書管理装置において、文字列索引内部に論理構造に関する情報を含めないことにより、文字列索引のサイズ縮小を可能とする効果を有する。更に文書の特定の要素内容の追加、変更、削除を行なう際に、処理量が大幅に軽減されるという効果を有する。

【0158】また、ノードの論理構造位置を特定するためのIDをパス名称IDとパス階層IDの2つに分けて管理しているため、論理構造が複雑かつ膨大になった場合でも、構造を特定するためのIDの総数を少なく押さえることを可能とする効果を有する。

【0159】また、各パス名称IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス名称ID照合テーブルや、各パス階層IDが検索条件の構造指定の範囲内にあるかどうかの情報が格納されるパス階層ID照合テーブルを作成し、構造照合処理を行なうことにより、中間ノード以下を指定した高速な検索を実現するという効果を有する。

【0160】なお、上述したように従来の技術では検索範囲として中間ノード以下を指定した場合、たとえ同一の親ノードを持つ同一タグ名を有するノードでも異なる文脈識別子が割り当てられるため、検索条件に該当するか否かをチェックする為のOR検索が必要となり、検索時間が大きくなるという課題を有していたが、本発明は、同一の親ノードを持つ同一タグ名を有するノードがたとえ複数存在しても、同一の識別子を付与するために、OR検索が不要となり、検索時間が短縮できるという効果を有する。

【0161】また、Mixed Contentにまたがる文字連鎖に対して拡張文字列索引を作成することによって、Mixed Contentにまたがる文字列を検索対象とすること、およびMixed Contentである要素を指定した検索を可能とする効果を有する。

【0162】また、あらかじめ設定されたタグ名の要素に対して数値型索引を作成することにより、要素内容を数値データとして扱った数値範囲の検索を可能とする効果を有する。

【図面の簡単な説明】

【図1】本発明の実施の形態1における構造化文書管理装置の構成図

【図2】本発明の実施の形態1における構造化文書の一例を示す図

【図3】本発明の実施の形態1における構造を解析した結果の木構造の一例を示す図

【図4】本発明の実施の形態1における構造情報作成手段の処理手順を示す図

【図5】本発明の実施の形態1における名称IDを割当てた例を示す図

【図6】本発明の実施の形態1におけるパス名称インデックスの一例を示す図

【図7】本発明の実施の形態1におけるパス階層インデックスの一例を示す図

【図8】本発明の実施の形態1における要素管理テーブルの一例を示す図

【図9】本発明の実施の形態1における文字列索引作成手段の処理手順を示す図

【図10】本発明の実施の形態1における文字列索引の一例を示す図

【図11】本発明の実施の形態1における検索条件解析手段の処理手順を示す図

【図12】本発明の実施の形態1における文字列索引を用いた検索処理の詳細を示す図

【図13】本発明の実施の形態1における構造照合手段の処理手順を示す図

【図14】本発明の実施の形態1における構造照合処理の詳細を示す図

【図15】本発明の実施の形態1におけるノード群を追加した木構造の一例を示す図

【図16】本発明の実施の形態2における構造化文書管理装置の構成図

【図17】本発明の実施の形態2における構造条件解析手段の処理手順を示す図

【図18】本発明の実施の形態2におけるパス名称ID照合テーブルの一例を示す図

【図19】本発明の実施の形態2における構造照合手段の処理手順を示す図

【図20】本発明の実施の形態2における構造照合処理の詳細を示す図

【図21】本発明の実施の形態2における構造照合手段で、中間ノードを指定した場合の検索範囲に該当するノードを特定するための処理手順を示す図

【図22】本発明の実施の形態2におけるパス階層ID照合テーブルの一例を示す図

【図23】本発明の実施の形態3における文字列索引作成手段の処理手順を示す図

【図24】本発明の実施の形態3における拡張文字列索引の一例を示す図

【図25】本発明の実施の形態4における構造化文書管理装置の構成図

【図26】本発明の実施の形態4における構造化文書の一例を示す図

【図27】本発明の実施の形態4における数値型索引作成手段の処理手順を示す図

【図28】本発明の実施の形態4における数値型索引の一例を示す図

【図29】本発明の実施の形態5における構造化文書管理装置の構成図

【図30】本発明の実施の形態5における文字列検索部の処理手順を示す図

【図31】本発明の実施の形態5における要素管理テーブルの一例を示す図

【図32】本発明の実施の形態5における要素管理テーブルの一例を示す図

【図33】従来の技術における文書登録システムの構成を示す図

【図34】従来の技術における構造インデックスの生成過程を示す図

【図35】従来の技術における文字列インデックスの例を示した図

【図36】従来の技術における構造インデックスの更新

方法を示した図

【符号の説明】

- 1 0 1 … 端末
- 1 0 2 … 構造化文書入力手段
- 1 0 3 … 検索条件入力手段
- 1 0 4 … 結果表示手段
- 1 0 5 … 検索エンジン
- 1 0 6 … データ格納部
- 1 0 7 … 構造解析手段
- 1 0 8 … 構造情報作成手段
- 1 0 9 … 文字列索引作成手段
- 1 1 0 … 検索条件解析手段
- 1 1 1 … 文字列索引検索手段
- 1 1 2 … 構造照合手段
- 1 1 3 … 結果作成手段
- 1 1 4 … 構造解析済みデータ格納手段
- 1 1 5 … 要素管理テーブル格納手段

- 1 1 6 … パス名称インデックス格納手段
- 1 1 7 … パス階層インデックス格納手段
- 1 1 8 … 名称IDテーブル格納手段
- 1 1 9 … 文字列索引格納手段
- 1 2 0 … 実体データ格納手段
- 1 2 1 … 一覧データ格納手段
- 1 6 0 1 … パス名称ID照合テーブル格納手段
- 1 6 0 2 … パス階層ID照合テーブル格納手段
- 2 4 0 1 … 数値型索引作成手段
- 2 4 0 2 … 数値型索引検索手段
- 2 4 0 3 … 数値型設定格納手段
- 2 4 0 4 … 数値型索引格納手段
- 3 0 0 1 … 構造化文書登録部
- 3 0 0 2 … 文字列索引作成部
- 3 0 0 3 … 文字列検索部
- 3 0 0 4 … 結果表示部

【図2】

【図5】

【図32】

```

<論文>
<書誌>
  <タイトル>構造化文書管理</タイトル>
  <著者>… </著者>
  <著者>… </著者>
  <日付>… </日付>
</書誌>
<本文>
  <章>
    <章タイトル>登録方法</章タイトル>
    <段落>登録方法においては、まず… </段落>
    <節>
      <節タイトル>文書入力手段</節タイトル>
      <段落>まず、構造化文書の入力は… </段落>
      <段落>実体管理においては、まず… </段落>
    </節>
    <節>
      <節タイトル>文書格納手段</節タイトル>
      <段落>文書格納の際には… </段落>
    </節>
  </章>
</本文>
</論文>
  
```

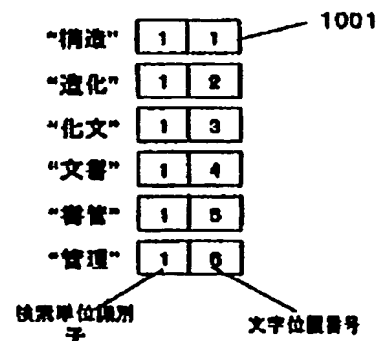
名称ID	名称	検索単位 識別子	文書番号	名称ID
T1	論文	1	1	T3
T2	書誌	2	1	T4
T3	タイトル	3	1	T4
T4	著者	4	1	T5
T5	日付	5	1	T8
T6	本文	6	1	T9
T7	章	7	1	T11
T8	章タイトル	8	1	T9
T9	段落	9	1	T9
T10	節	10	1	T11
T11	節タイトル	11	1	T9
⋮	⋮	⋮	⋮	⋮

名称IDテーブル

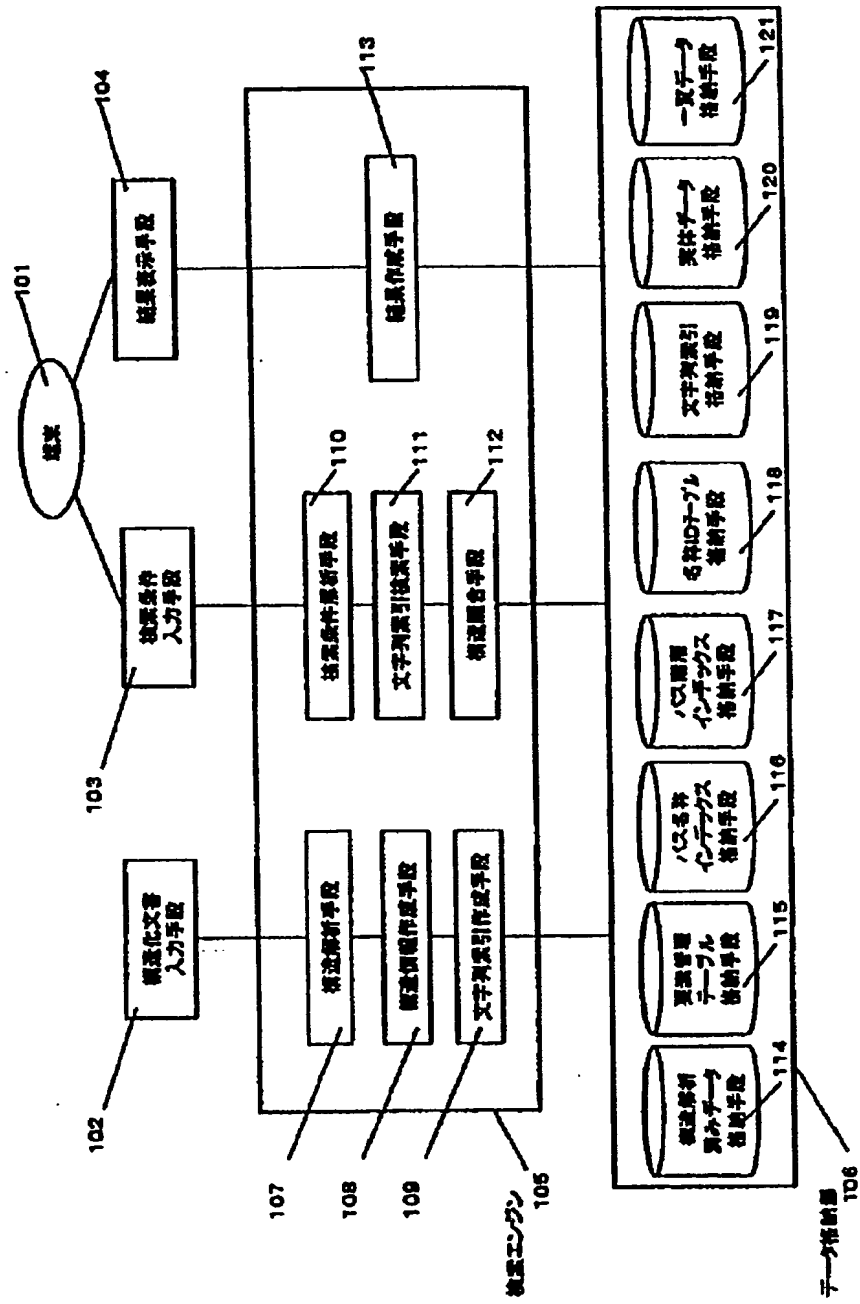
要素管理テーブル

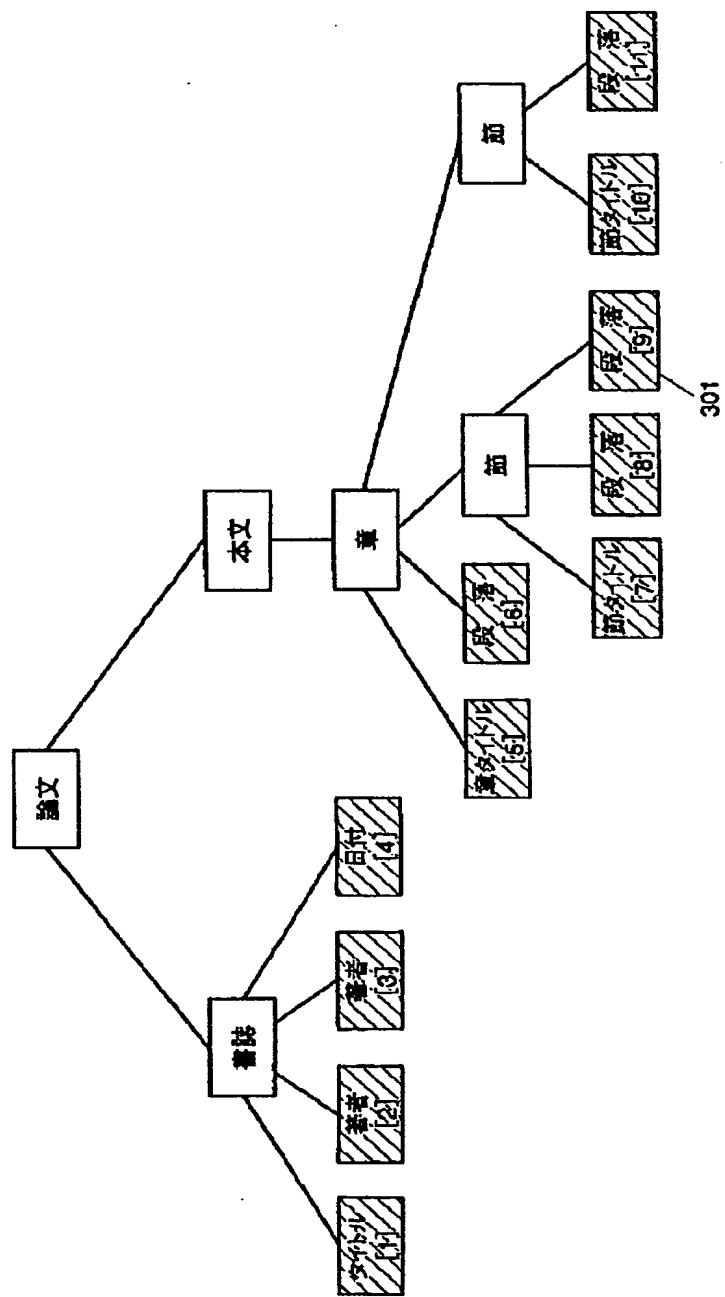
【図10】

<タイトル> 構造化文書管理 </タイトル>



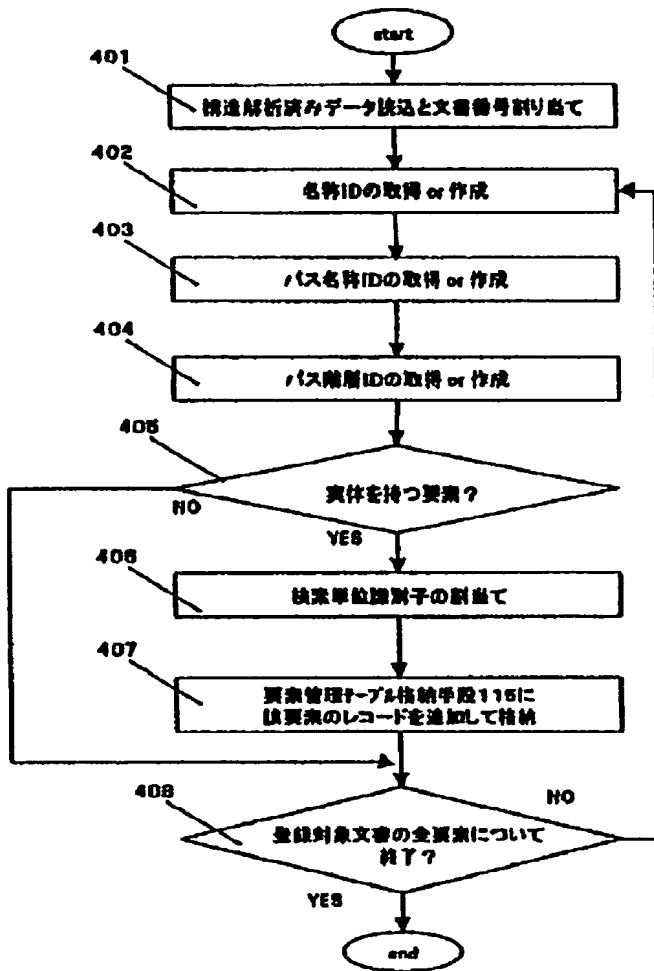
【図1】





【図 3】

【図4】



【図18】

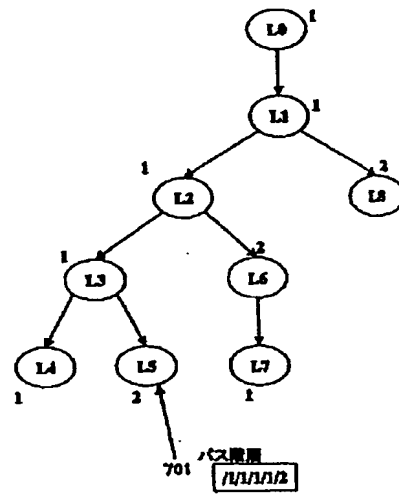
バス名称ID	整合フラグ
N0	2
N1	2
N2	2
N3	2
N4	2
N5	2
N6	1
N7	1
N8	1
N9	1
N10	1
N11	1

バス名称ID整合テーブル

バス階層ID	整合フラグ
L0	2
L1	2
L2	1
L3	1
L4	1
L5	1
L6	1
L7	1
L8	2

バス階層ID整合テーブル

【図7】

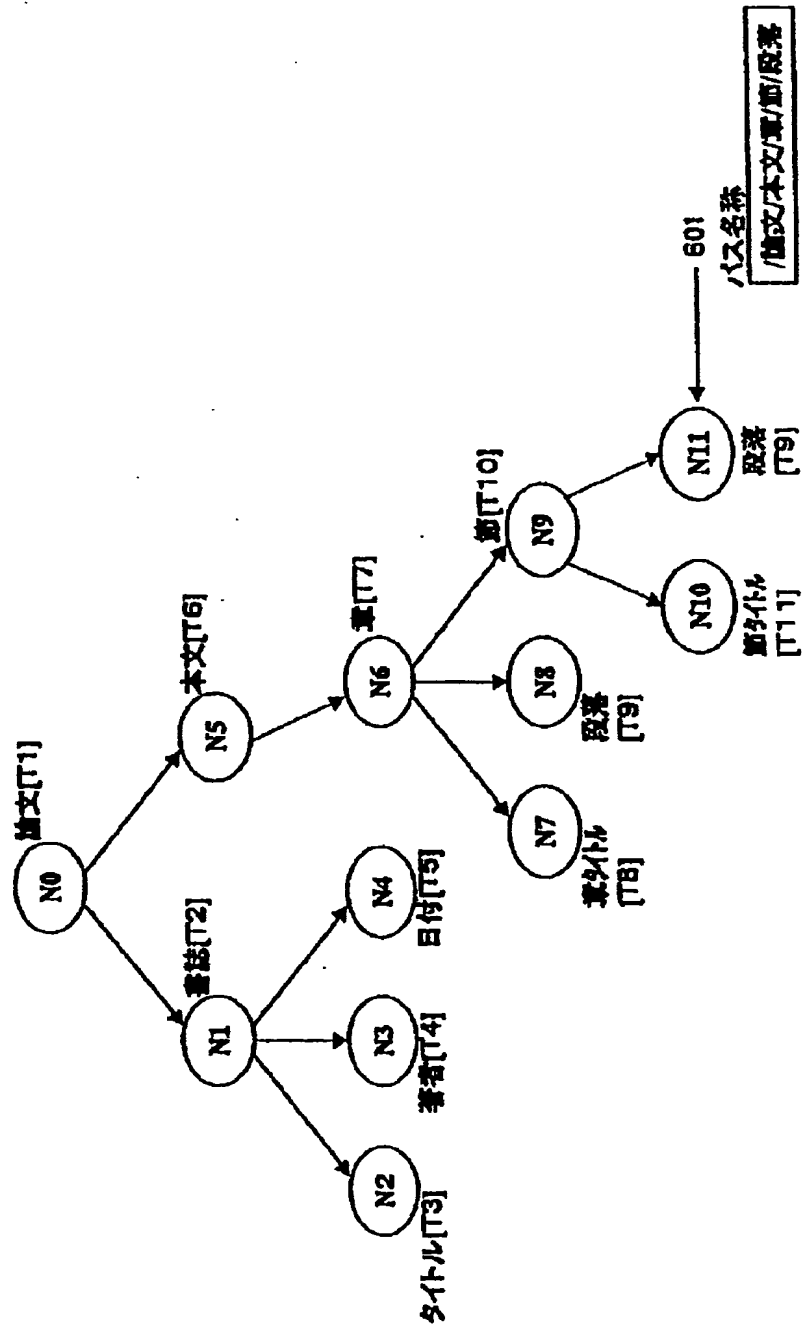


【図8】

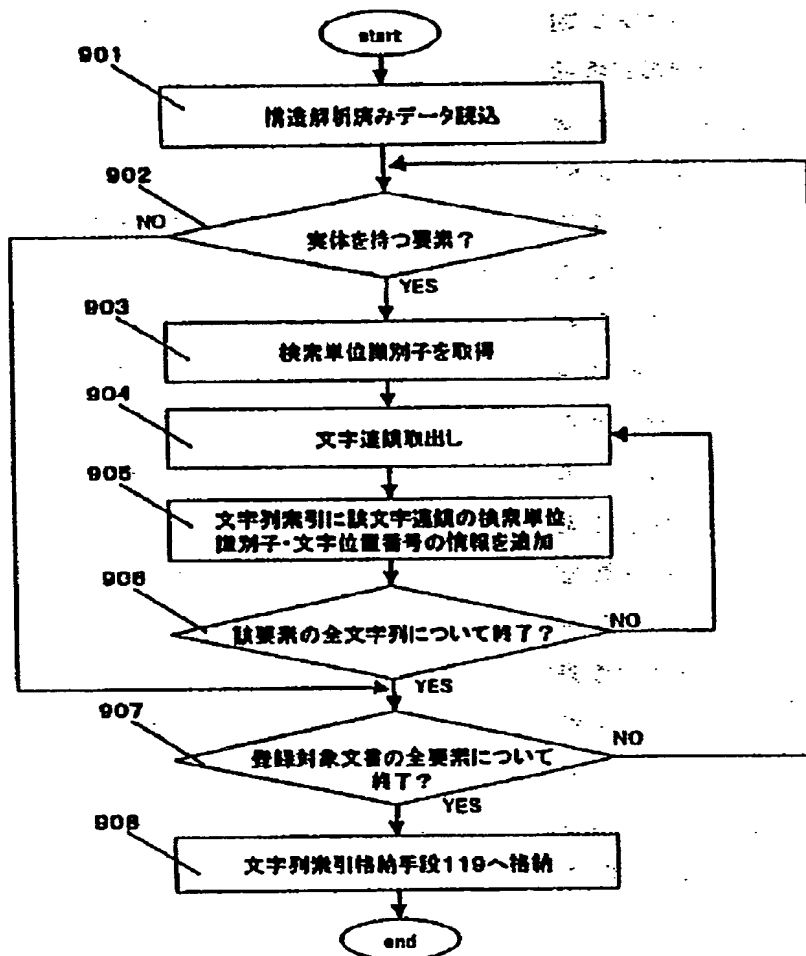
検索単位 識別子	文書番号	バス名称ID	バス階層ID	要素ID
1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L8	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
9	1	N11	L5	T9
10	1	N10	L7	T11
11	1	N11	L7	T9
...

要素管理テーブル

【図6】



【図9】



【図28】

2710

検索単位 識別子	数値データ
201	000001600

2720

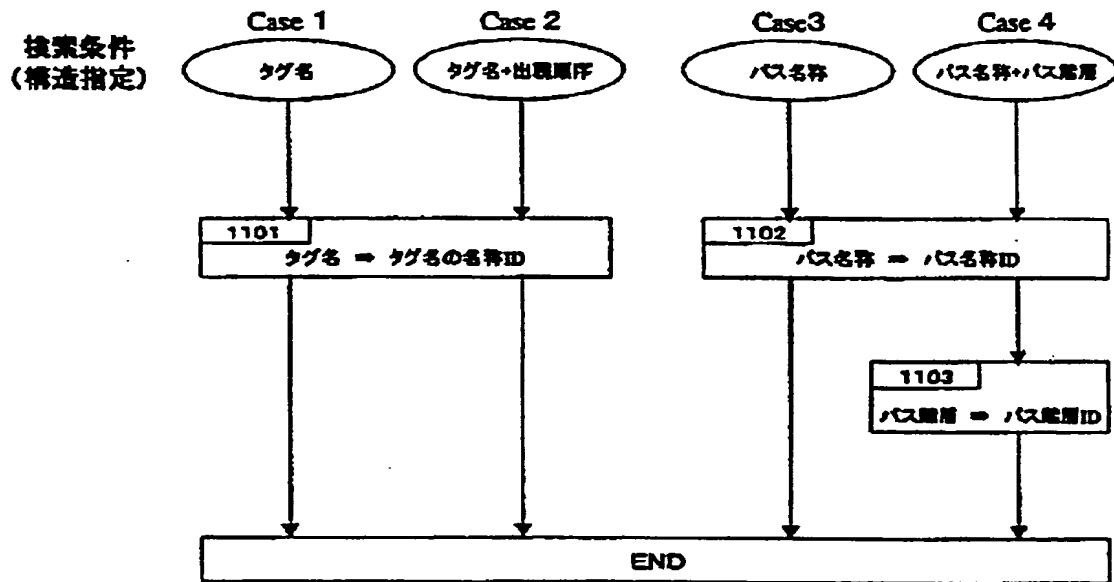
検索単位 識別子	数値データ
25	000002998
54	000001598
105	000019800
201	000001600
231	000000670
545	000001600
1352	000000800

【図31】

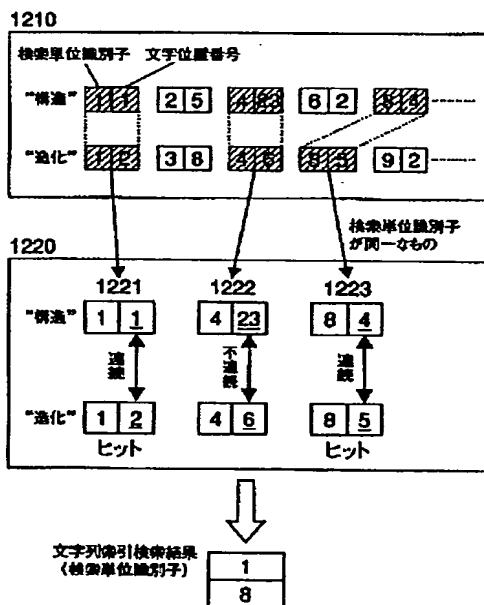
検索単位 識別子	文書番号	パス名称ID	パス属性ID
1	1	N2	L2
2	1	N3	L2
3	1	N3	L3
4	1	N4	L2
5	1	N7	L3
6	1	N8	L3
7	1	N10	L4
8	1	N11	L4
9	1	N11	L5
10	1	N10	L7
11	1	N11	L7
...

要素管理テーブル

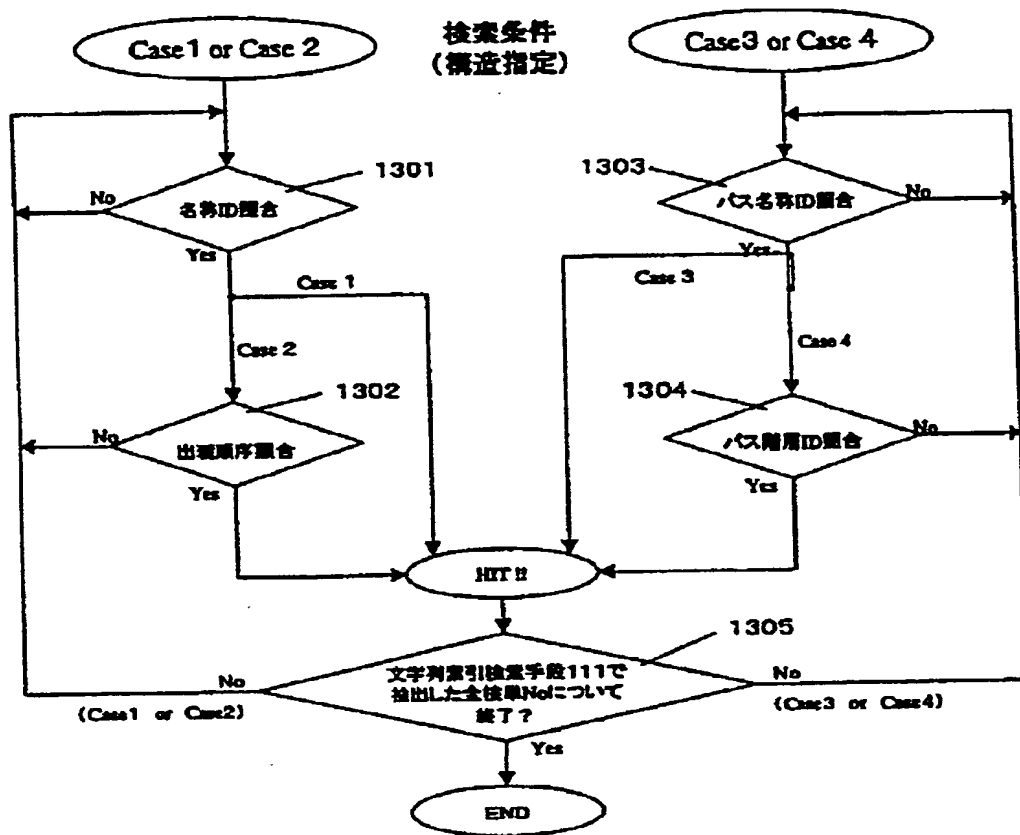
【図11】



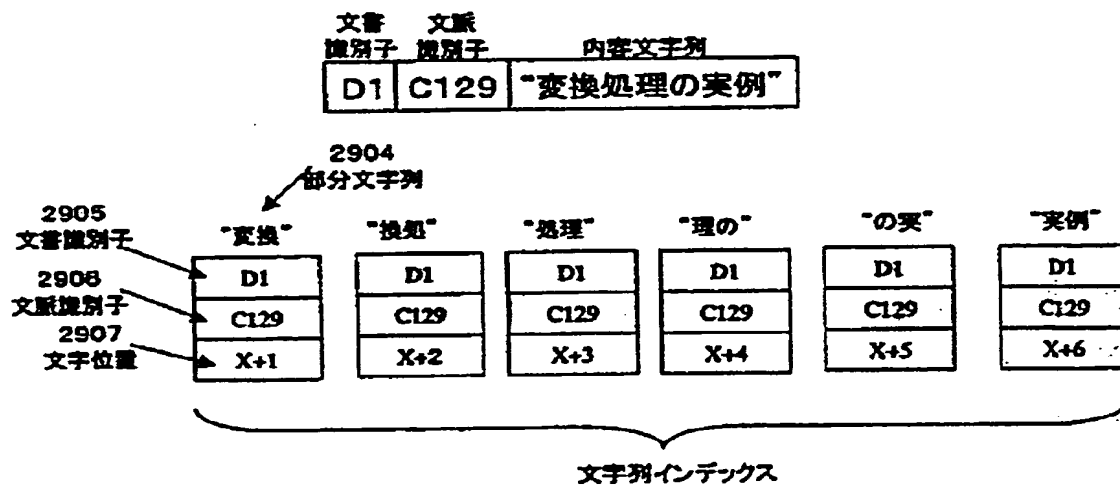
【図12】



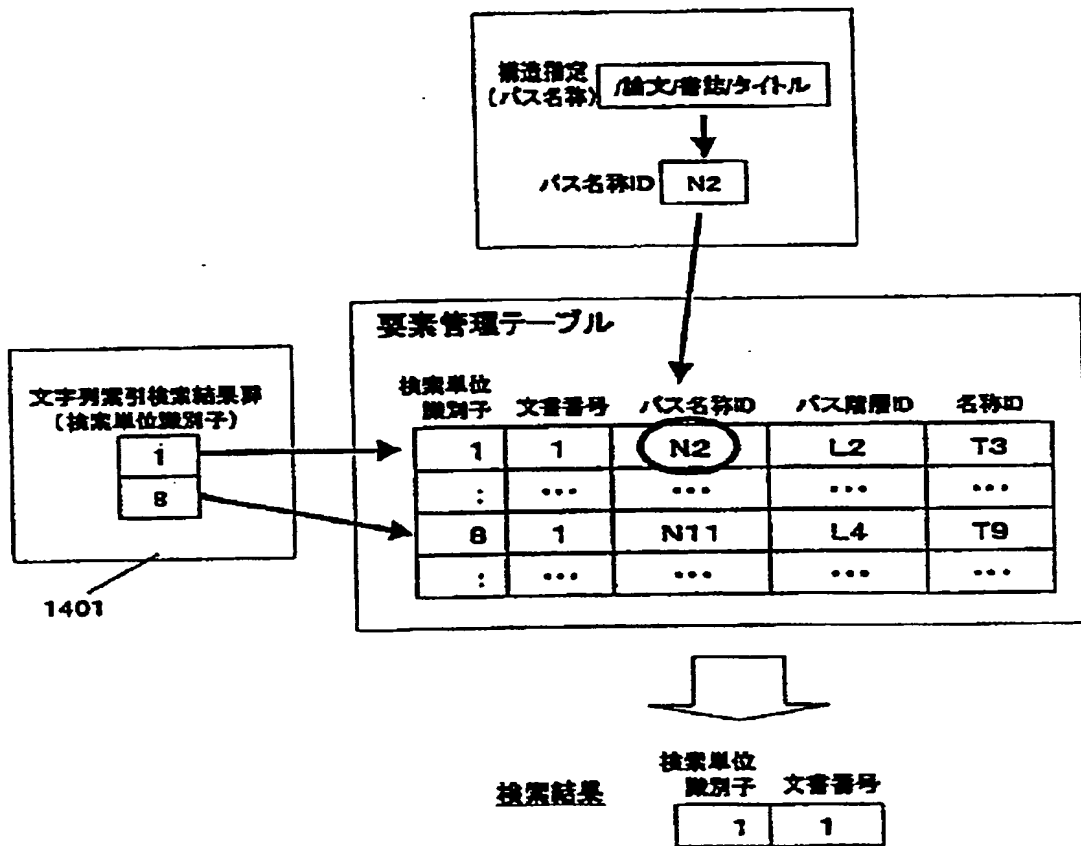
【図13】



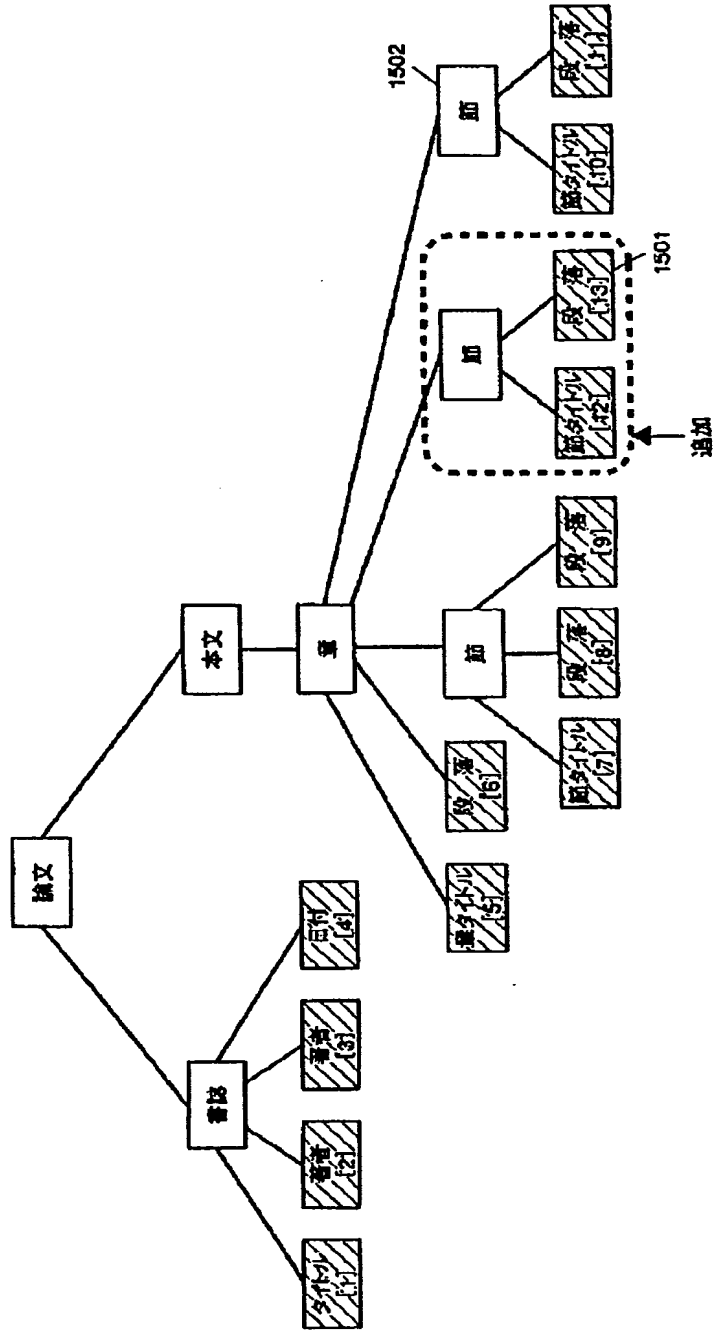
【図35】



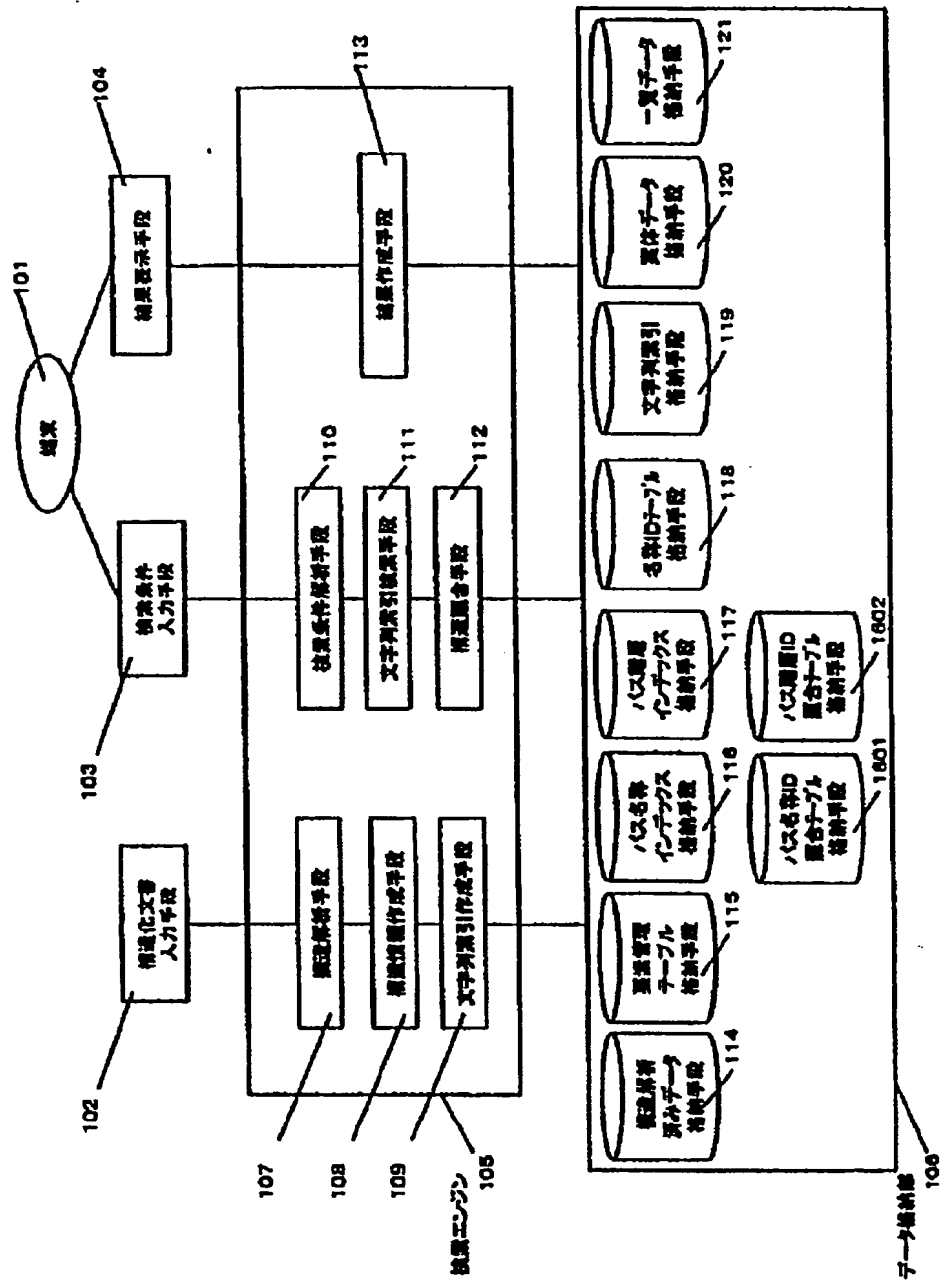
【図14】



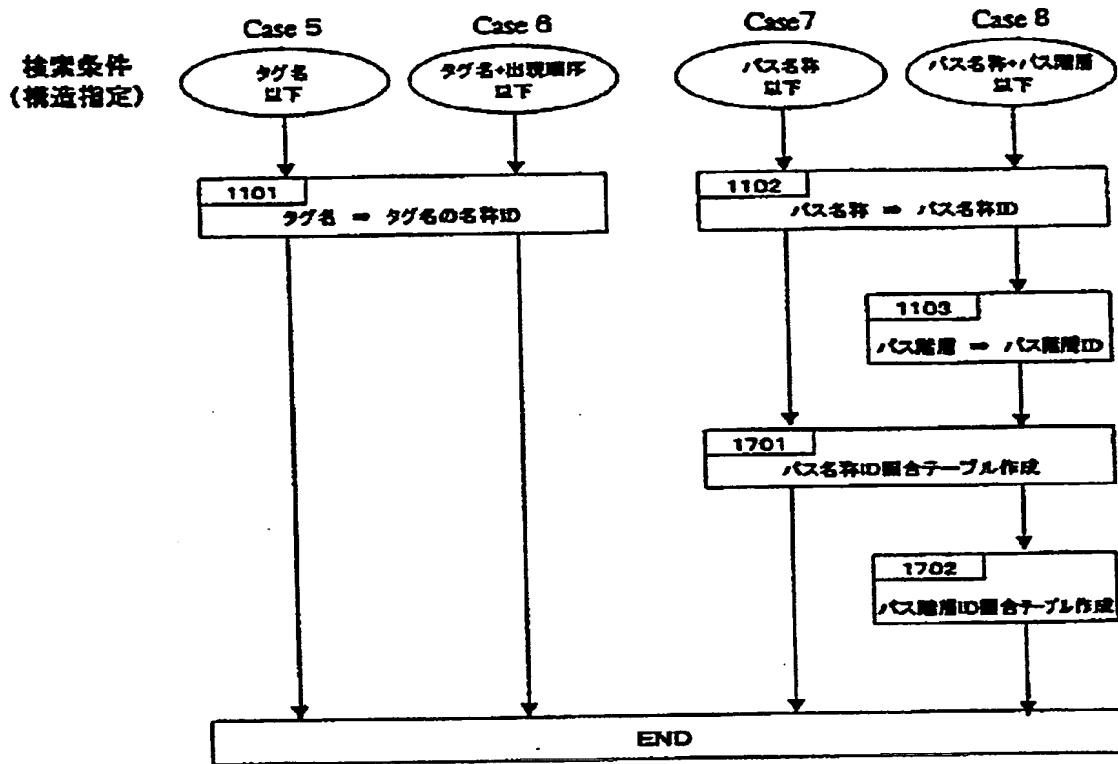
【例 15】



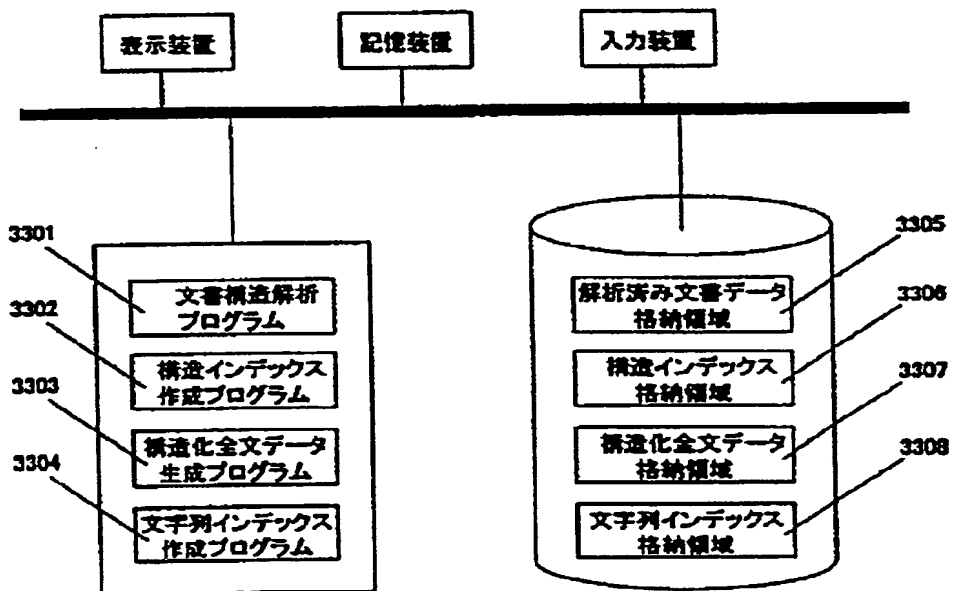
-31-



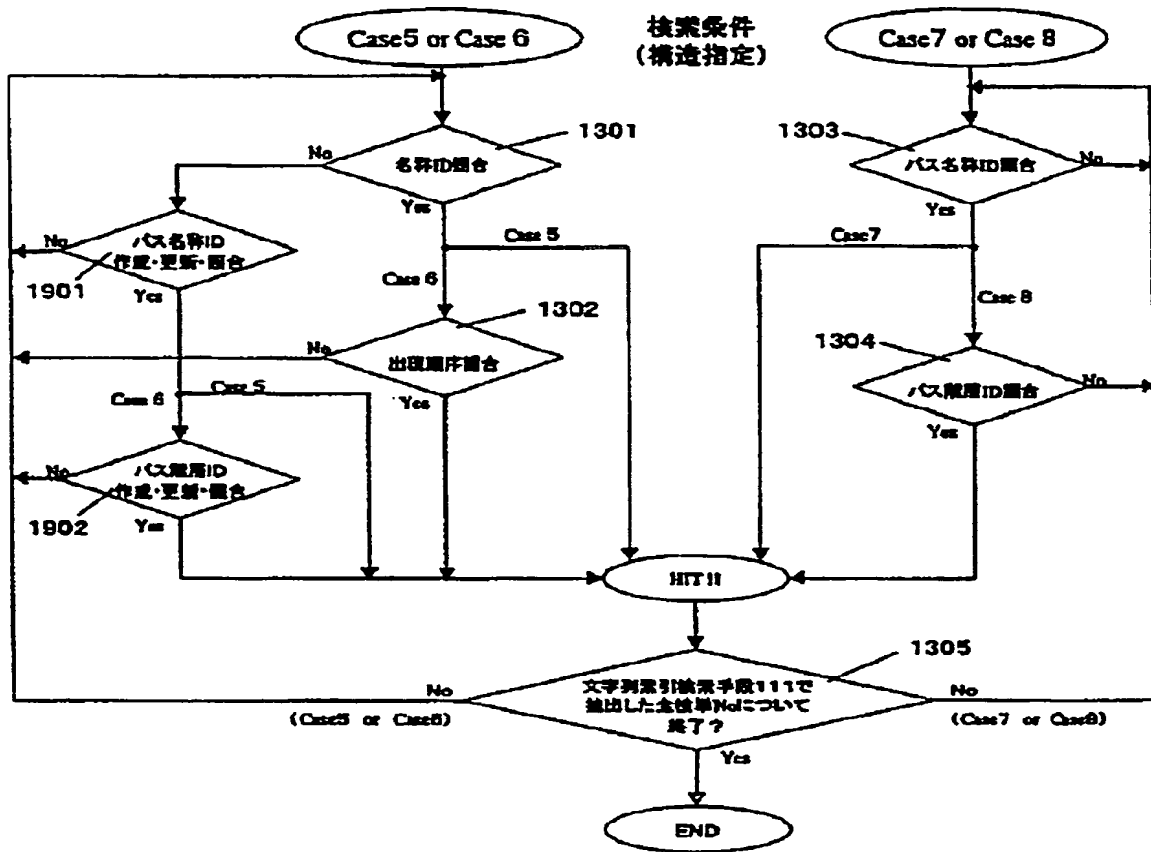
【図17】



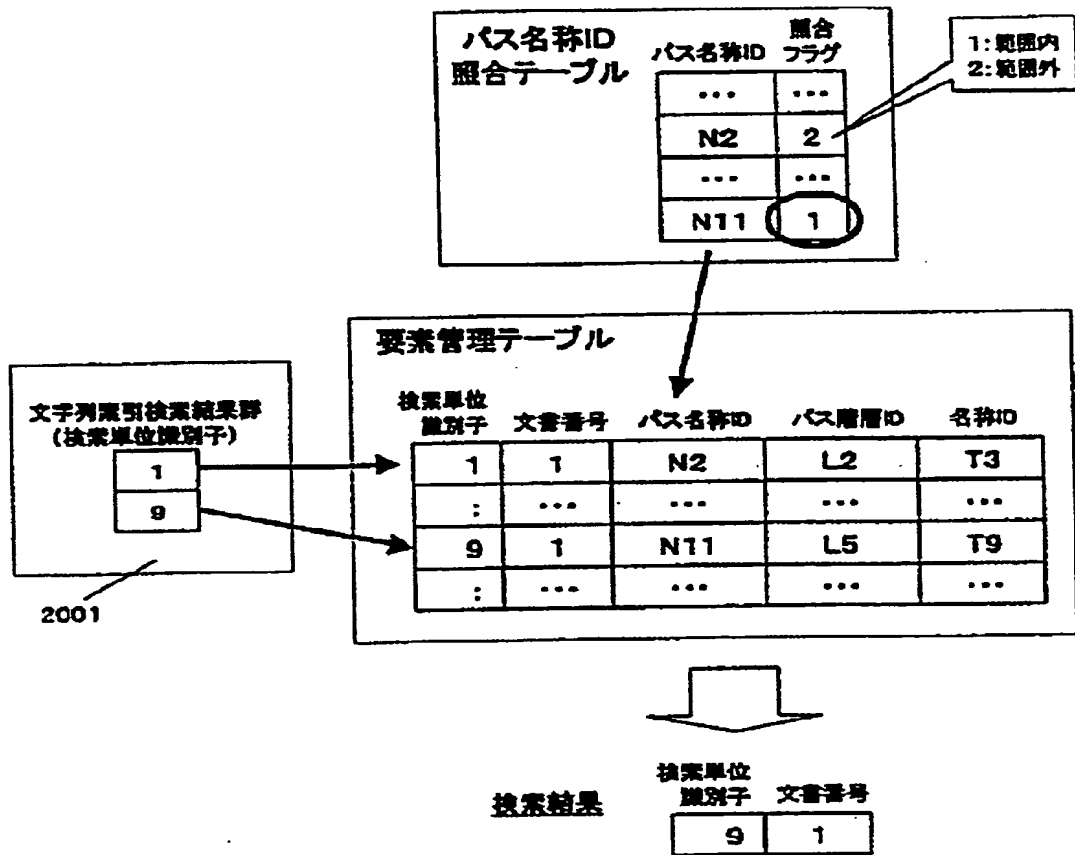
【図33】



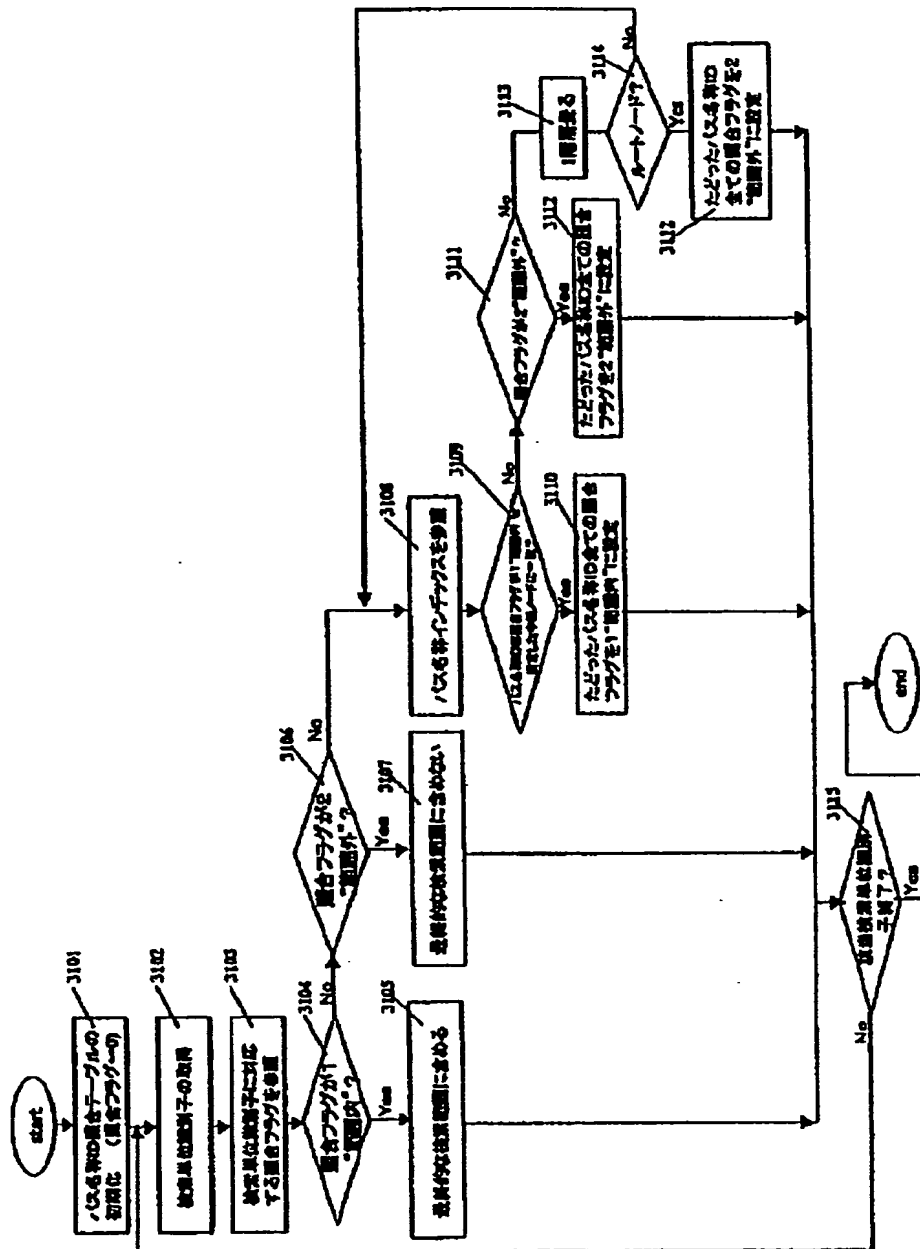
【図19】



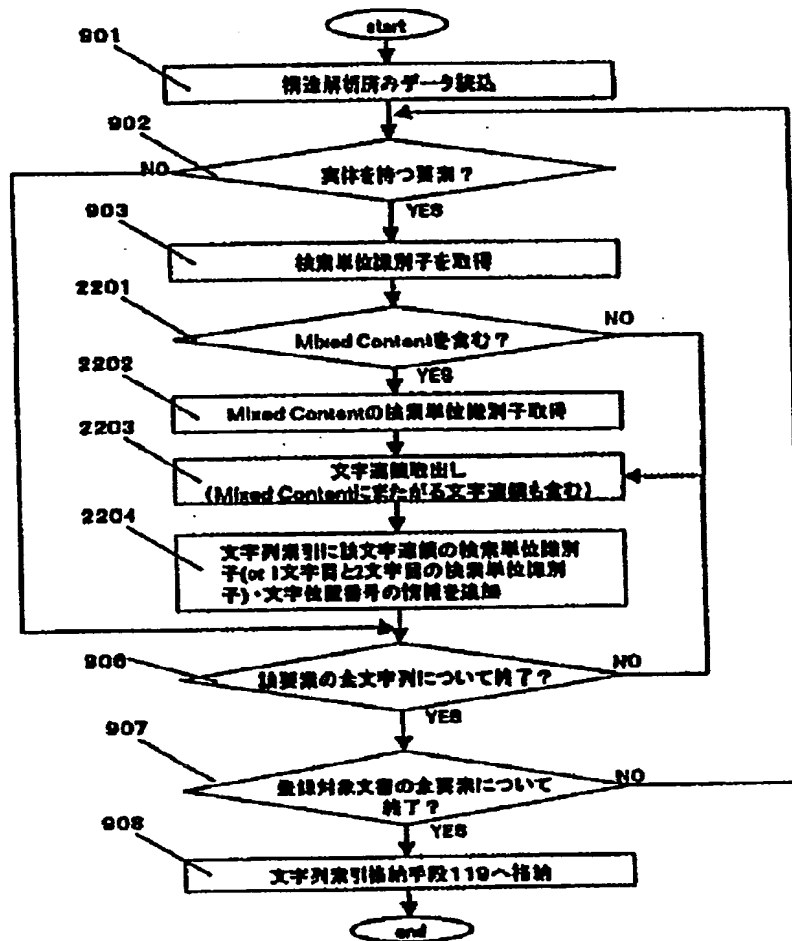
【図20】



【図21】

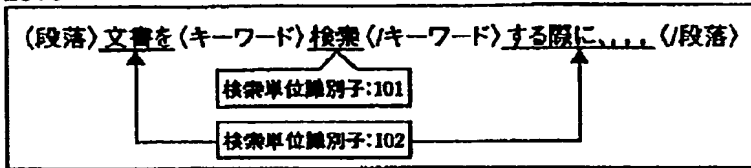


【図23】

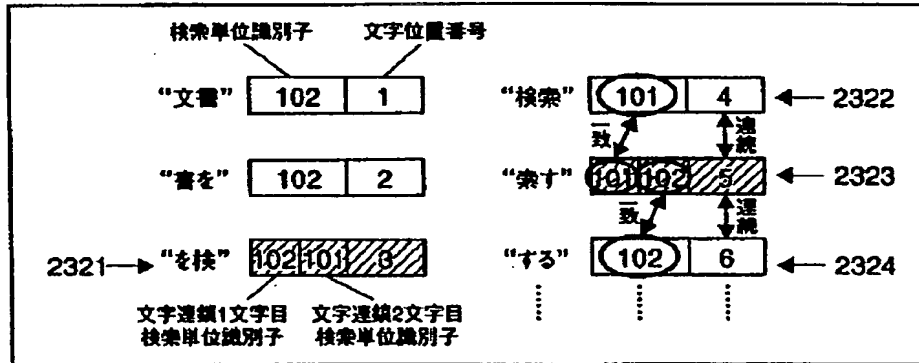


【図24】

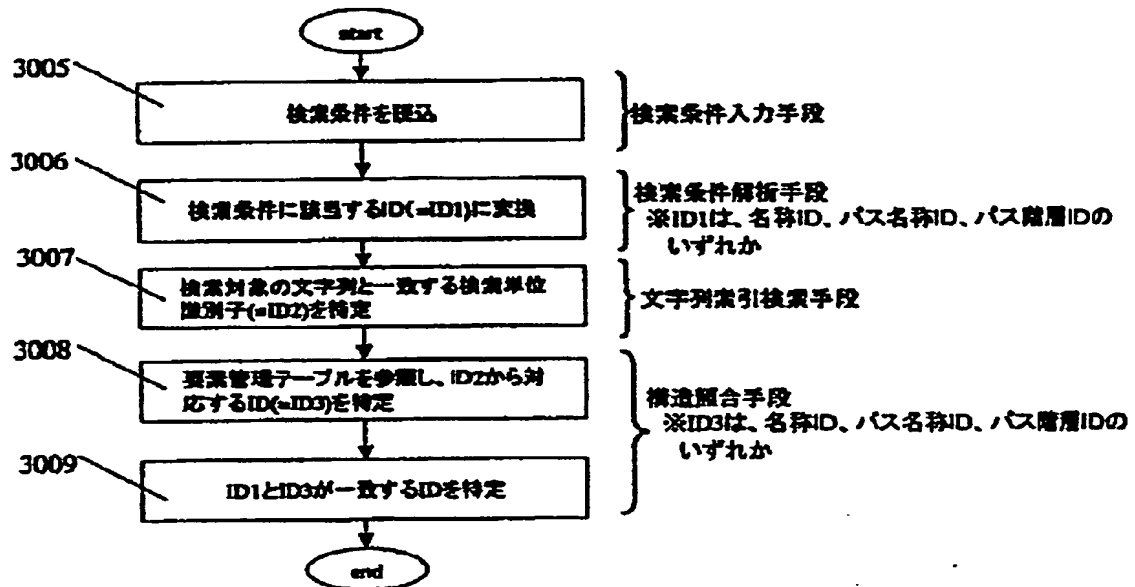
2310



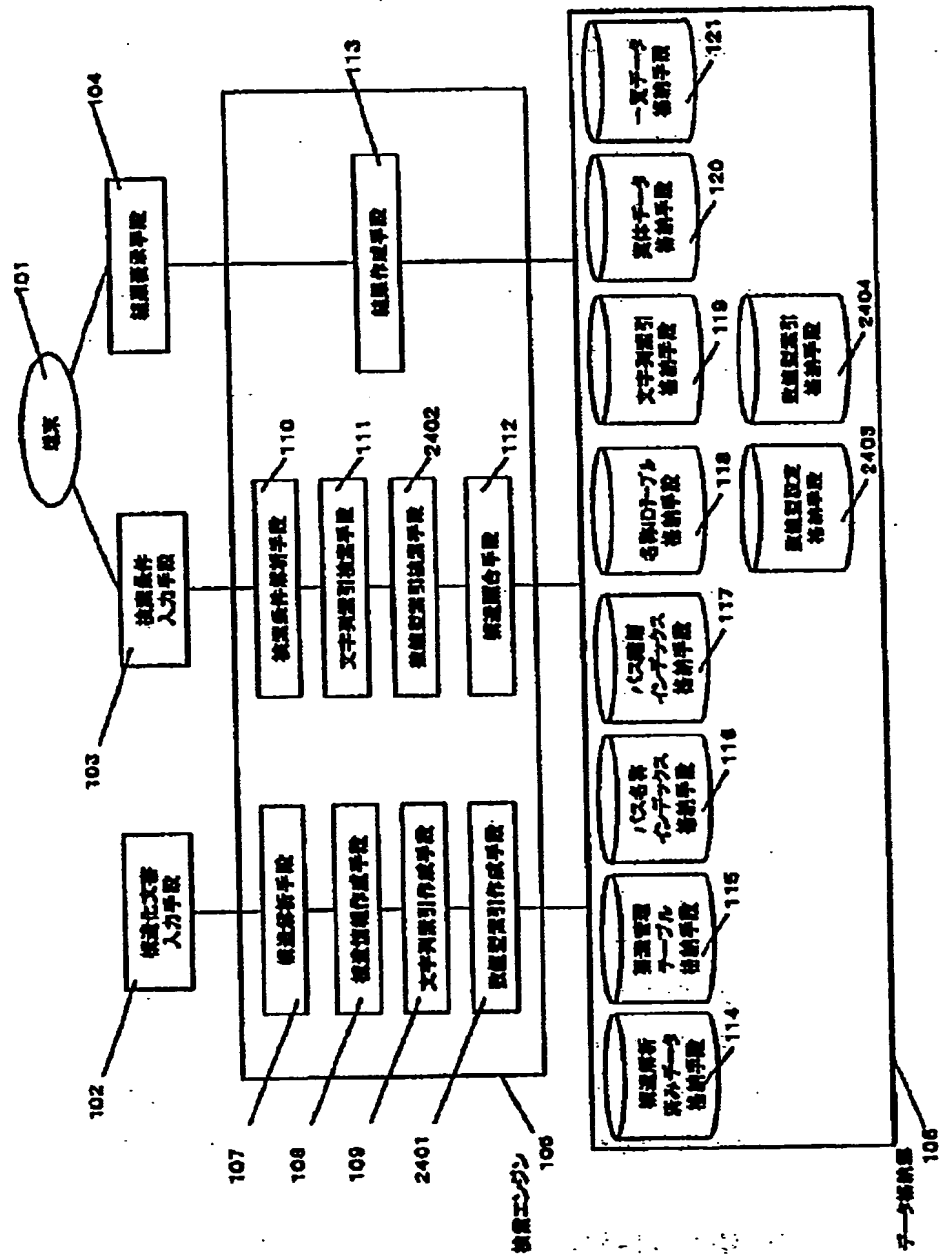
2320



【図30】



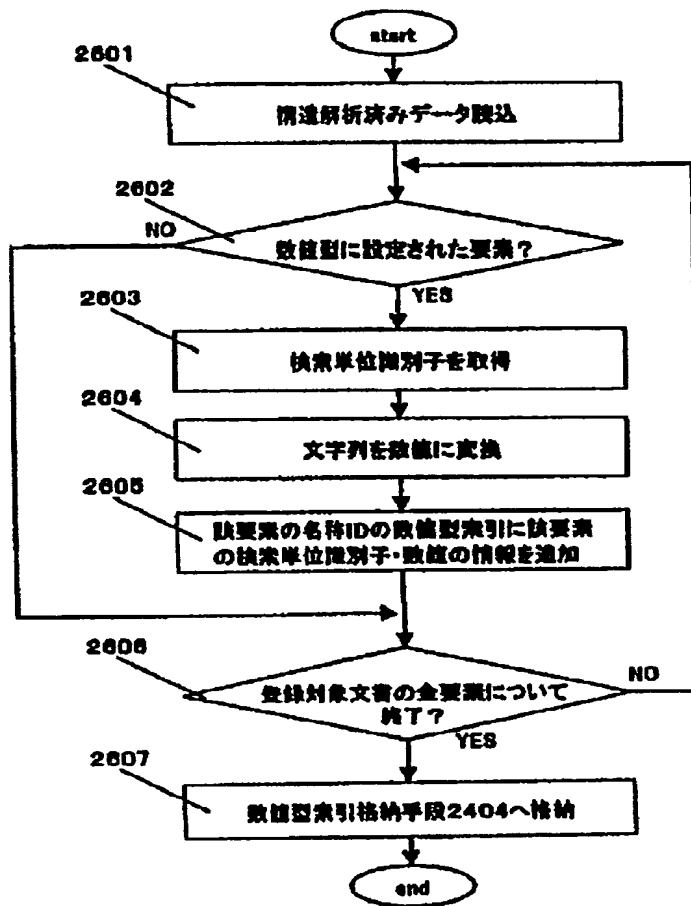
【図25】



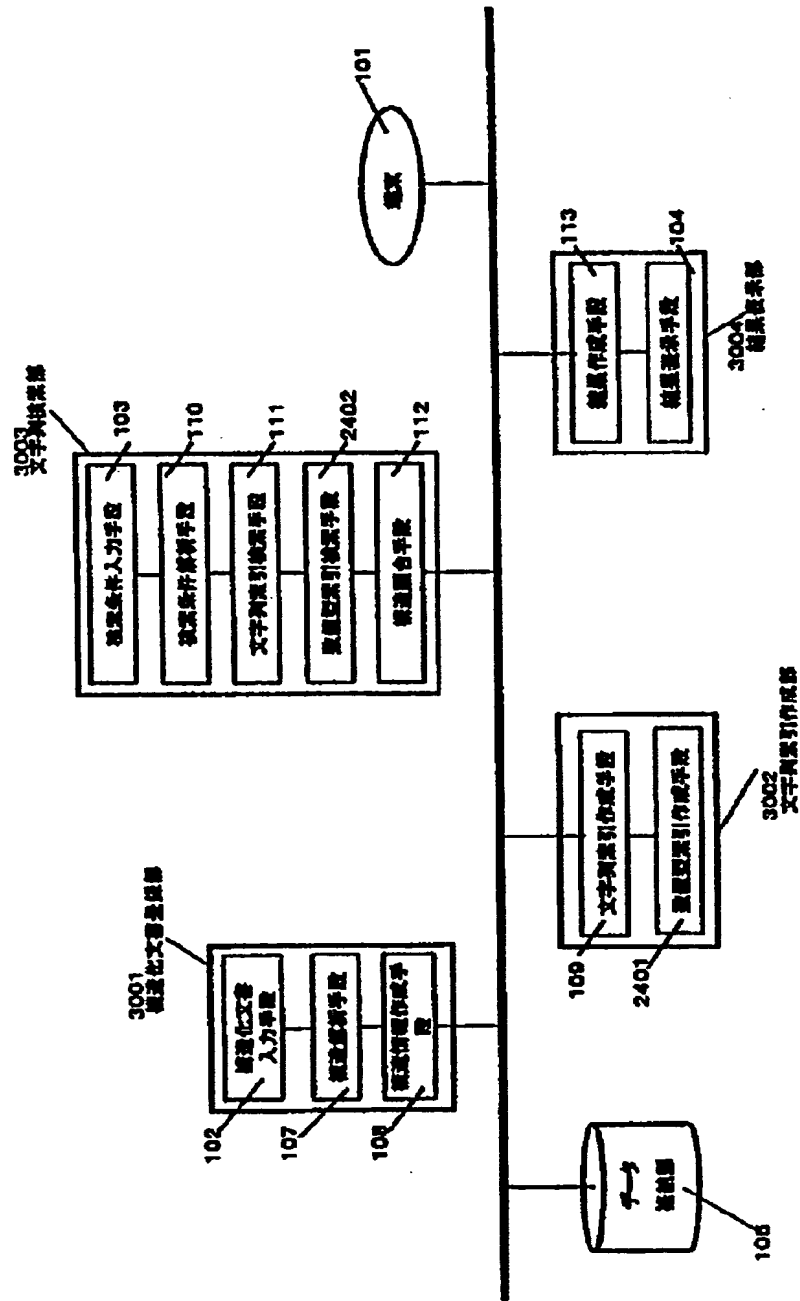
【図26】

<書籍>
<書誌>
 <タイトル>...</タイトル>
 <著者>...</著者>
 <出版社>...</出版社>
 <価格>1600円</価格> ← 2501
</書誌>
<本文>
 <章>
 <章タイトル>...</章タイトル>
 <段落>...</段落>
 <節>
 <節タイトル>...</節タイトル>
 <段落>...</段落>
 :
 :
 :
 :
 </節>
</章>
</本文>
</書籍>

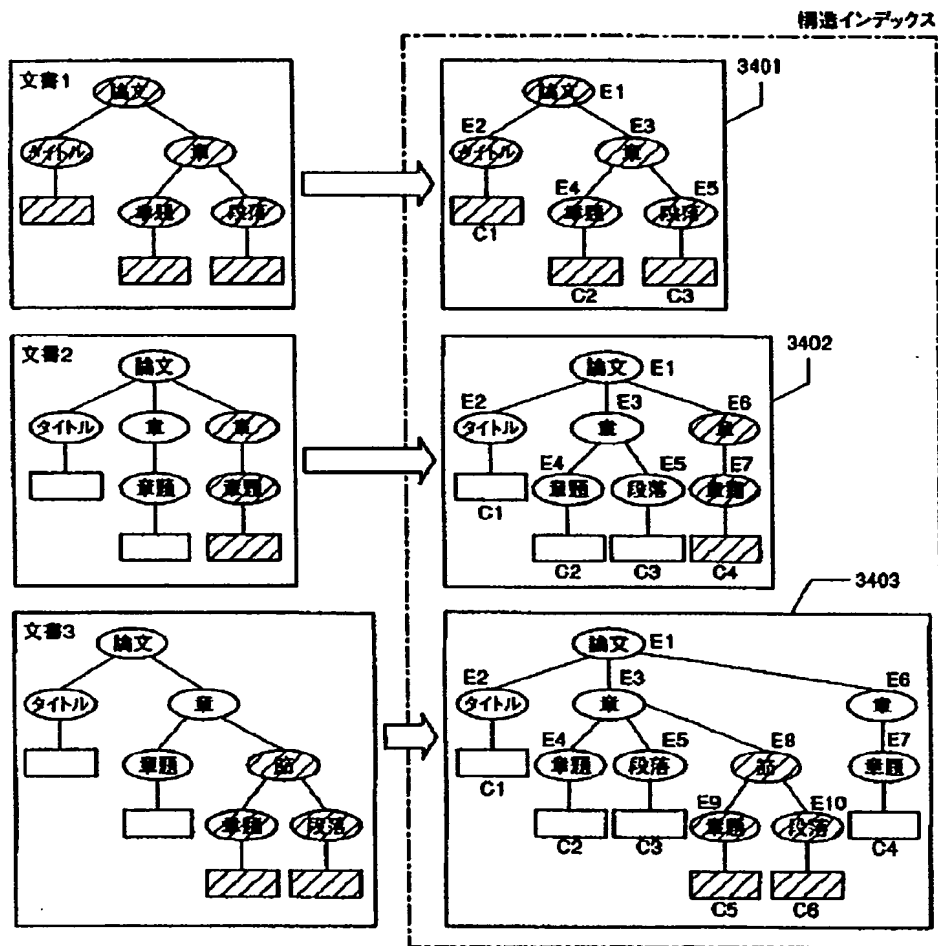
【図27】



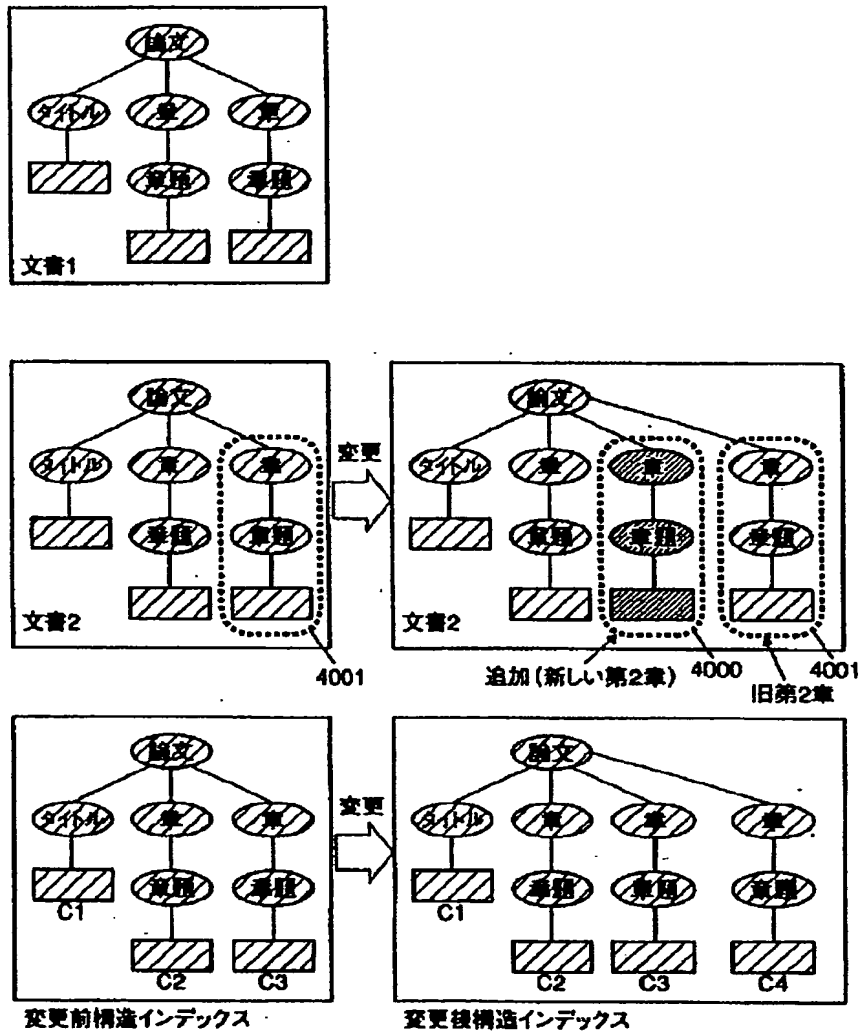
【图29】



【図34】



【図36】



フロントページの続き

(72)発明者 鶴林 健
大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72)発明者 片山 修
大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72)発明者 中井 信一
大阪府門真市大字門真1006番地 松下電器
産業株式会社内

Fターム(参考) 5B075 ND35 NK43

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record.

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☒ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER: _____**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.